CANCER TREATMENT OPTIMIZATION

A Thesis
Presented to
The Academic Faculty

by

Kyungduck Cha

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology
April 2008

CANCER TREATMENT OPTIMIZATION

Approved by:

Professor Eva K. Lee, Advisor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Earl R. Barnes
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Ellis L. Johnson
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Renato D.C. Monteiro
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Nolan E. Hertel
G. W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Date Approved: March 28, 2008

To my parents, wife, and son.

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere appreciation to my advisor, Dr. Eva Lee for her invaluable guidance, financial support, and understanding throughout my Ph.D program of study. Her inspiration provided me many ideas in tackling with my thesis problems and her enthusiasm encouraged me to produce better works.

I would also like to thank Dr. Earl Barnes, Dr. Ellis Johnson, Dr. Renato D.C. Monteiro, and Dr. Nolan Hertel for their willingness to serve on my dissertation committee and for their valuable feedbacks.

I also thank my friends, faculty, and staff, at Georgia Tech, who have contributed to this thesis in various ways. Specially, thank you to the members at the Center for Operations Research in Medicine and HealthCare for their friendship.

I really appreciate my parents for the endless support and the sacrificial love. To my son, Jiho Cha, thank you for growing up healthy. You are the one that remind me the reason of the life. To my wife, Woonghwi Lee, thank you for staying with me. You are my big tree where I take a good rest. It would be impossible for me to complete the Ph.D without your support and love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This dissertation investigates optimization approaches applied to radiation therapy in cancer treatment. Since cancerous cells are surrounded by critical organs and normal tissues, there is conflicting objectives in the treatment design of providing sufficient radiation dose to tumor region, while avoiding normal healthy cells. In general, the goal of radiation therapy is to conform the spatial distribution of the prescribed dose to the tumor volume while minimizing the dose to the surrounding normal structures. A recent advanced technology, using multi-leaf collimator integrated into linear accelerator, provides much better opportunities to achieve this goal: the radiotherapy based on non-uniform radiation beams intensities is called Intensity-Modulated Radiation Therapy.

In this thesis, we offer a quadratic mixed integer programming approach to determine optimal beam orientations and beamlets intensity simultaneously. The problems generated from real patient cases are large-scale dense instances due to the physics of dose contributions from beamlets to volume elements. The research highlights computational techniques to improve solution times for these intractable instances. Furthermore, results from this research will provide plans that are clinically acceptable and superior in plan quality, thus directly improve the curity rate and lower the normal tissue complication for cancer patients.

CHAPTER I

OVERVIEW: OPTIMIZATION IN INTENSITY MODULATED RADIATION THERAPY

Eva K. Lee, PhD[1,2], Joseph O Deasy, PhD[3]

[1]Center for Operations Research in Medicine

School of Industrial and Systems Engineering

Georgia Institute of Technology

Atlanta GA 30332-0205, USA

[2]Winship Cancer Institute and Department of Radiation Oncology

Emory University School of Medicine

[3]Division of Bioinformatics and Outcomes Research

Department of Radiation Oncology

Washington University School of Medicine

4921 Parkview Place, LL, Campus Box 8224

St. Louis, MO 63110

Abstract

In this chapter, we provide an overview and some computational challenges in intensity modulated radiation therapy. This chapter is an article written by Lee and Deasy for SIAM Views and News. Here, experience with a mixed-integer programming treatment planning model is described. The MIP model allows simultaneous optimization over the space of beamlet intensity weights and beam and couch angles. The model uses two classes of decision variables to capture the beam configuration and intensities simultaneously. Binary (0/1) variables are used to capture "on" or

"off" or "yes" or "no" decisions for each field, and nonnegative continuous variables are used to represent intensities of beamlets. Binary and continuous variables are also used for each voxel to capture dose level and dose deviation from target bounds. The treatment planning model was designed to explicitly incorporate the following planning constraints: (a) upper/lower/mean dose-based constraints, (b) dose-volume and equivalent-uniform-dose constraints for critical structures, (c) homogeneity constraints (underdose/overdose) for the planning target volume (PTV), (d) coverage constraints for PTV, and (e) maximum number of beams allowed. Results of applying the MIP Model to a patient case are presented. Brief discussions of recent linear programming and nonlinear programming treatment planning models are also described, as is an MIP approach for direct aperture optimization.

Keywords Optimization, Intensity-Modulated Radiation Therapy, Mixed Integer Programming, Beam Angle and Intensity Map Optimization

1.1 Introduction

Every year over 1.4 million new cancer cases are diagnosed [95] in the United States, and over half of the patients receive radiation treatment at some point during the course of their disease. The key to the effectiveness of radiation therapy for the treatment of cancer lies both in the fact that the repair mechanisms for cancerous cells are less efficient than that of normal cells, and the ability to deliver higher doses to the target volume using "cross-fired" radiation beam. Thus, a dose of radiation sufficient to kill cancerous cells may not be lethal for nearby healthy tissue. Nevertheless, avoiding or minimizing radiation exposure to healthy tissue is extremely important.

Using multiple beams of radiation from multiple directions to cross-fire at the tumor volume provides a method to keep radiation exposure to normal tissue at relatively low levels, while dose to tumor cells is elevated. The crux of the treatment planning process involves designing beam profiles (i.e., a collection of beams) that delivers a sterilizing dose of radiation to the tumor volume, while dose levels to critical normal tissues are kept below established tolerance levels. Often, one attempts to design a plan for which the prescription dose isodose surface conforms to the geometric shape of the specified tumor volume. [48, 109] (The term *prescription dose* typically refers to the minimum dose desired to be delivered to the tumor volume; it is generally physician specified.)

Linear accelerators (LINAC) are common beam delivery units used for external beam radiotherapy. The table on which the patient lies and the beam delivery mechanism for the LINAC rotate about separate orthogonal axes, providing the ability to adjust the angle and entry point of radiation fields used during treatments. Each field is further defined by a bank of multi-leaf collimators (MLC), small metallic leaves located inside the LINAC treatment unit. These leaves can be opened or closed, and used to shape the radiation beam as it exits the machine. Figure 1 shows a linear accelerator.

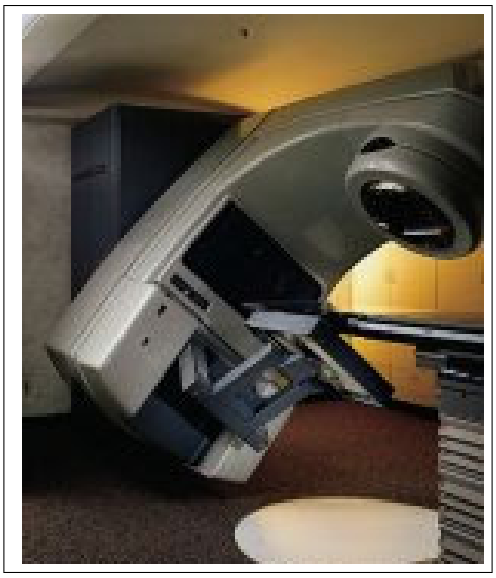


Figure 1: A Linear accelerator used for external beam radiotherapy treatment

Intensity-modulated radiation therapy (IMRT) is an important recent advance in radiation therapy [110]. In conventional radiotherapy treatment, the planning process consists of determining a set of external beams that meet, as best as possible, the clinical dose distribution criteria. In many cases, significant compromises to critical structure function have to be made to enable a tumoricidal dose to be delivered to the targets. In IMRT, the radiation fluence is varied across the beam, which allows a higher degree of conformation to the tumor than previously possible and allows concave isodose profiles to be generated, which may block, for example, dose to critical structure anterior or posterior to the target from that view. Specifically, not only is the shape of the beam controlled, but combinations of open and closed multileaf collimators modulate the intensity as well. For this reason, IMRT provides improved delivery control over conventional treatment. Indeed, it provides an unprecedented capability to dynamically vary the dose to accommodate the shape of the tumor from different angles, and to spare normal tissues and organs-at-risk (OAR) that may be potentially harmed during treatment.

Due to the complexity of the beam intensity profile associated with IMRT, there

has been a tremendous research effort among medical physicists and radiation oncologists related to IMRT treatment planning and delivery, and there remain many opportunities for computational advances, particularly in treatment design. A computer-driven optimization algorithm must be used to determine the beam fluences (intensity maps) that provide the best compromise between target underdosing, target overdosing and critical structure overdosing. The textbook by Webb [110] has a good list of references for IMRT optimization.

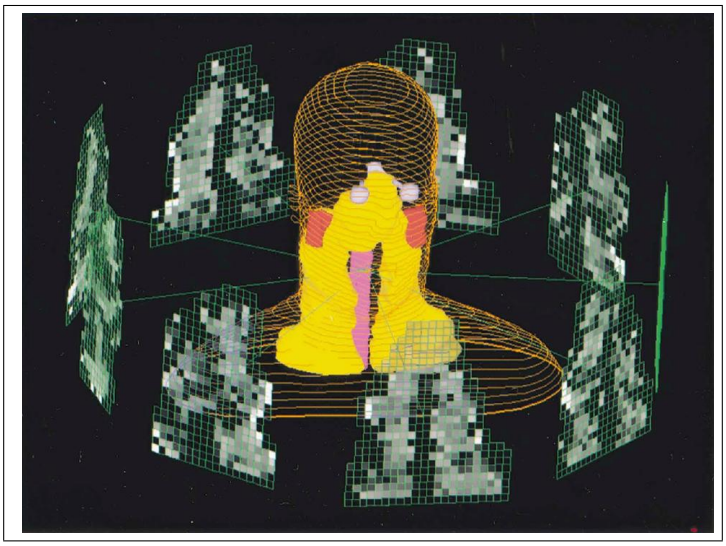


Figure 2: The treatment of a head-and-neck case via IMRT. Shown is a 3D view of the patient, the planning tumor volume (PTV), yellow; the spinal cord, pink; and the parotid glands, red. The 9 beams, shown with gray levels, reflect the modulated radiation intensity. (Use with permission from [39])

In Sections 2.1, 2.2, we describe the treatment planning problem for IMRT, and discuss relevant input data and the dose matrix. In Section 2.3, we discuss our experience of a mixed integer programming treatment planning model. The mixed integer programming model allows one to simultaneously incorporate dose coverage, underdose, overdose, homogeneity and conformity criteria on the tumor volume; dose volume restrictions on the critical structures (how much volume can receive more than a specified dose); and physical constraints on the total number of beams. Section 2.4 describes briefly the associated clinical results, and Section 2.5 provides a very

brief discussion of current mathematical programming approaches. Summary and discussion is presented in Section 3.

1.2 Intensity-Modulated Radiation Therapy Treatment Planning

Treatment planning in intensity modulated radiation therapy consists of a sequence of steps:

- Acquiring a 3D image of the affected region
- Delineating target volumes and healthy anatomical structures
- Selecting the appropriate radiation source and energy
- Selecting a set of beam angles for use in treatment
- Computing dose from each beam
- Performing intensity map optimization for the selected beams
- Developing optimal collimator sequences for actual delivery

These steps can be performed sequentially, or some can be combined together, resulting in complex numerical problems. In the sections presented herein, much of the description follows our recent work on mixed integer programming in this area [55, 56]. Very brief discussions on linear programming and nonlinear programming approaches are included.

1.2.1 Input Data and Dose Calculation

Image Acquisition and Segmentation. The planning process begins when the patient is diagnosed with a tumor mass and radiation is selected as part of the treatment regime. A 3D image, or volumetric study-set, of the affected region, which contains the tumor mass and the surrounding areas, is acquired via computed tomography (CT) scans. These CT data are used for treatment planning, and electron density

information derived from them are used in the photon dose calculations. Additionally, magnetic resonance imaging (MRI) scans may be acquired, fused with the CT volumetric study-set, and used to more accurately identify the location and extent of some tumors – especially those in the brain. Based on these scans, the physician outlines the tumor, and also outlines anatomic structures that need to be held to a low dose during treatment.

It is common practice to identify three "volumes" associated with the tumor. The gross tumor volume (GTV) represents the volume that encompasses the imageable or palpable macroscopic disease; that is, the disease that can be detected and localized by the oncologist. The clinical target volume (CTV) expands the GTV to include regions of suspected microscopic disease. The delineation of the CTV depends heavily on a *priori* knowledge of the behavior of a given tumor type. For a given GTV, tumor histologic features, and patient type, a set of probabilities exist (imperfectly known) that the tumor will, or will not, extend microscopically into a given regional organ or lymph node. However, accurate specific data are usually not available to the radiation oncologist, only general principles are known. A more quantitative and consistent definition of the CTV is an important need. The planning target volume (PTV) includes additional margins for anatomical and patient setup uncertainties related to organ and patient movement over time. All volumetric data is discretized into voxels (point representations of volume properties) at a granularity that is conducive both to generating a realistic model and to ensuring that the resulting treatment planning instances are tractable (i.e., capable of being solved in a reasonable amount of computational time for practical clinical usage).

Dose Calculation Radiation dose, measured in Gray (Gy), is energy (Joules) deposited locally per unit mass (kg). Fluence for external beam photon radiation is defined mathematically by the number of photon crossings per surface area. Dose tends to be proportional to fluence, but is also influenced by photons and electrons

scattered in the patient's tissues as well as the incident energy and media involved.

The calculation of the dose distribution associated with IMRT delivery is a critical aspect of the IMRT optimization and delivery processes. The calculated dose distribution from each candidate set of plan parameters is evaluated at each iteration or at the end of the optimization process, and the objective function values (costs or scores) for the iterative optimization are typically obtained by analysis of the dose distribution. For most systems, after the fluence-optimized plan is obtained, another dose calculation/optimization procedure, called leaf sequencing, is performed which first breaks the beams up into machine-deliverable multileaf sequencing steps, and then includes a final dose calculation step based on the details of the multileaf field shapes.

One of the most commonly used IMRT dose calculation algorithms involves a simple pencil beam method and is usually part of a broader class of correction-based dose-calculation algorithms [2, 66]. While these models offer significant speed advantages for use in the optimization code, they have varying limitations in accuracy.

In contrast, convolution/superposition, energy deposition kernel-based approaches can take into account beam energy, geometry, beam modifiers, patient contour, and electron density distribution [1, 8, 15, 67, 70]. Both the convolution method and the Monte Carlo method compute the dose per unit energy fluence (or fluence) incident on the patient.

Although it is clear that improved dose-calculation accuracy afforded by the convolution-type calculations may be important for IMRT, the long calculation times make this difficult.

Recently, significant progress has been made in the development of Monte Carlo calculation algorithms for photon beams, which simulate particle tracks individually, that are fast enough to compete with other current methods [17, 40, 90, 91]. In several situations, the Monte Carlo method is likely to be even more accurate than

the convolution method [111]. For example, multiple scatter (second and higher order scatter) may be perturbed near the surface of a patient and the Monte Carlo method may be able to account for this as long as the number of simulated particles is sufficient. Direct Monte Carlo simulation may be the only option for achieving accurate dose computations in these complex situations. However, the application of Monte Carlo methods to optimization for IMRT is an area that requires much more work before relevant results will be available.

Access and usage of realistic radiotherapy data can be facilitated by using an open-source toolbox, developed by Deasy et al [29], which enables users to import clinical plan data into Matlab for viewing and manipulation, and furthermore includes tools to generate the dose influence matrices.

1.2.2 Treatment Planning Strategies

In a strategy known as forward treatment planning, the beam geometry (beam orientation, shape, modifier, beam weights, etc.) is first defined, followed by calculation of the 3D dose distribution. After qualitative review of the dose distribution by the treatment planner and/or radiation oncologist, plan improvement is often attempted by modifying the initial geometry (e.g., changing the beam weights and/or modifiers, adding another beam), to improve the target dose coverage and/or decrease the dose in the organs at risk. This forward planning process is repeated until a satisfactory plan is generated. As one can imagine, this is a time consuming approach to treatment planning.

In newer inverse treatment planning, the focus is on the desired outcome (e.g., a specified dose distribution or tumor control probability (TCP) and normal tissue complication probability (NTCP)) rather than on how the outcome is achieved. The user of the system specifies the goals; the computer (optimization system) then adjusts the beam parameters (mainly the intensities) iteratively in an attempt to achieve the

desired outcome. After review of the computer optimized dose distribution, some modification of the desired outcome and adjustment of the relative importance of each end point might be needed if the physician is not satisfied with the dose to the target volume or organs-at-risk (OARs).

Clearly, optimization is a classical inverse planning approach: constraints and an objective function are utilized to guide the optimization solver to select a plan with pre-specified clinical properties. Beginning with the work of Bahr et al [5] in the late 60's, a number of research articles, authored primarily by medical researchers, discussed the use of mathematical programming and other optimization techniques in conventional external beam radiation treatment planning [27, 50, 51, 52, 54, 87, 96, 100, 117].

Much of IMRT treatment planning research has focused on the determination of the fluence map[11, 13, 16, 18, 19, 24, 28, 36, 39, 41, 42, 43, 71, 72, 109, 113]; that is, the radiation intensity or beam weights associated to each of the small beamlets of a selected radiation field/beam. However, the determination of beam angles, shapes, modifiers, couch positions and radiation energy to be used are best modelled using discrete variables.

At present, most IMRT optimization systems use dose-based and/or dose-volume-based criteria. One method commonly used to create dose-based and dose-volume objective functions involves minimizing the variance of the dose relative to the prescribed dose for the target volumes or dose limits for the organs at risk. This type of objective function has been used for traditional radiation therapy treatment optimization for the past several decades [97]. Variance is defined as the sum of the squares of the differences between the calculated dose and the prescribed dose or dose limit. Thus, a typical dose-based or dose-volume-based objective function is the sum of the variance terms representing each anatomic structure multiplied with

appropriate penalty factors (i.e., importance factors). Just as in conventional radiation therapy [23], the resulting unconstrained quadratic programming problem is often solved via the gradient method [96, 113], although the inclusion of dose-volume constraints makes the problem non-convex [30].

Within the optimization community, linear programming and nonlinear programming have been used to determine the optimal intensity map [86, 92], while mixed integer programming has been introduced to simultaneously determine the optimal beam angles and beam intensities [55, 56], and in finding optimal apertures for radiation delivery [83]. Below, we describe the MIP models formulated for simultaneous beam angle and intensity map optimization, closely following the presentation in Lee et al [55, 56]. Results from a patient case will be briefly summarized. We then briefly describe linear and nonlinear programming approaches by others. Besides mathematical programming approaches, heuristic approaches — such as simulated annealing and genetic algorithms — have been commonly used for radiation therapy treatment optimization.

1.2.3 Mixed Integer Programming Treatment Planning Models

The treatment planning models in [55, 56] use two classes of decision variables to capture the beam configuration and intensities simultaneously: Binary (0/1) variables are used to capture "on" or "off" or "yes" or "no" decisions for each field, and nonnegative continuous variables are used to represent intensities of beamlets. Binary and continuous variables are also used for each voxel to capture dose level and dose deviation from target bounds. Below, we provide the mathematical description of the treatment planning models.

Let $\mathcal{B}$ denote the set of candidate beams (each with an associated beam angle), and let $\mathcal{N}_i$ denote the set of beamlets (discretized sub-beams – usually rectangular in cross-section – which comprise the beam) associated with beam $i \in \mathcal{B}$. Beamlets associated

with a beam can only be used when the beam is chosen to be "on." If a beam is on, the beamlets with positive dose intensity will contribute a certain amount of radiation dosage to each voxel in the target volume and other anatomical structures. Once the set of potential beamlet intensities is specified, the total radiation dose received at each voxel can be modelled. Let $w_{ij} \geq 0$ denote the intensity of beamlet j from beam i (in calibrated monitor units). Then the total radiation dose at a voxel P is given by

$$D_P(w) = \sum_{i \in \mathcal{B}} \sum_{j \in \mathcal{N}_i} D_{P,ij}\ w_{ij}, \tag{1}$$

where $D_{P,ij}$ denotes the dose per monitor unit intensity contribution to voxel P from beamlet j in beam i. Various dose constraints are involved in the design of treatment plans. Clinically prescribed lower and upper bounds, say L_P and U_P, for dose at voxel P are incorporated with (1) to form the basic dosimetric constraints:

$$\sum_{i \in \mathcal{B}} \sum_{j \in \mathcal{N}_i} D_{P,ij}\ w_{ij} \geq L_P \qquad \text{and} \qquad \sum_{i \in \mathcal{B}} \sum_{j \in \mathcal{N}_i} D_{P,ij}\ w_{ij} \leq U_P. \tag{2}$$

Our model also allows selection of optimal beam angles out of a collection of candidate beams. Thus, the resulting plan returns the optimal beam geometry as well as beam intensities.

Let x_i be a binary variable denoting the use or non-use of beam i. The following constraints limit the total number of beams used in the final plan and ensure that beamlet intensities are zero for beams not chosen:

$$\sum_{i \in \mathcal{B}} x_i \leq B_{\max} \qquad \text{and} \qquad w_{ij} \leq M_i x_i \tag{3}$$

Here, $B_{\max}$ is the maximum number of beams desired in an optimal plan, and M_i is a positive constant that can be chosen as the largest possible intensity emitted from beam i.

For each voxel in each anatomical structure, we associate one binary variable and one continuous variable to capture whether or not desired dose level is achieved and the deviation of received dose from desired dose. We also impose additional constraints into our treatment plan design, as discussed below.

Clinically, it may be desirable to incorporate coverage constraints within the model. For example, the clinicians may consider that it is acceptable if, say, 95% of the PTV receives the prescription dose, *PrDose*. Such a coverage requirement can be modelled as follows.

$$\sum_{i\in\mathcal{B}}\sum_{j\in\mathcal{N}_i} D_{P,ij}w_{ij} - r_P = PrDose,\ P \in PTV \tag{4}$$

$$r_P \leq D^{OD}_{PTV}v_P \tag{5}$$

$$r_P \geq D^{UD}_{PTV}(v_P - 1) \tag{6}$$

$$\sum_{P\in PTV} v_P \geq \alpha|PTV|. \tag{7}$$

Here, r_P is a real-valued variable that measures the discrepancy between prescription dose and actual dose; v_P is a 0/1 variable that captures whether voxel P is above or below the prescription dose bounds or not; α corresponds to the minimum percentage of coverage required (e.g., $\alpha = 0.95$); D^{OD}_{PTV} and D^{UD}_{PTV} are the maximum overdose and maximum underdose levels tolerated for tumor cells; and $|PTV|$ represents the total number of voxels used to represent the planning target volume. The values D^{OD}_{PTV} and D^{UD}_{PTV} can be chosen according to the homogeneity level desired by the clinician for the resulting plan. If r_P0, then voxel P receives sufficient radiation dose to cover the prescribed dose. In this case, $v_P = 1$ and the amount of radiation for voxel P above the prescribed dose is controlled by the maximum-allowed-overdose constant, D^{OD}_{PTV}. Similarly, when $r_P < 0$, voxel P is underdosed, and the amount of underdose is limited by D^{UD}_{PTV}. In this case, $v_P = 0$.

By design, constraints (5) and (6) serve two purposes: 1) they capture the number

of PTV voxels satisfying the prescription dose, and 2) they provide a means of controlling underdose, overdose, and dose homogeneity in the tumor. For the latter, the ratio $(PrDose + D^{OD}_{PTV})/(PrDose - D^{UD}_{PTV})$ can be viewed as an implied PTV homogeneity constraint associated with the model. Using a model with a smaller homogeneity constraint can be expected to result in a more homogeneous plan. Constraint (7) corresponds to the coverage level desired by the clinician.

Recently Equation (4) has been used to capture dose gradient fall-off when 100% tumor coverage is demanded. This was achieved by minimizing the dose surrounding the tumor region [54]. For IMRT planning optimization, it alone was used to model the deviation from prescribed dose for the PTV [19, 25, 113]. In these studies, a nonlinear objective function was formulated to steer the gradient-based optimization engine towards achieving the prescribed dose for the target volume; specifically, the objective was to minimize the sum of dose deviation across the target volume: $\|r\|_q = (\sum_P |r_P|^q)^{1/q}$ (with no imposed constraints). When $q = 2$, this is a least-squares problem.

It is desirable that dose received by radiation sensitive organs/tissues other than the tumor volume should be controlled to reduce the risk of injury. Thus, for other anatomical structures involved in the planning process, along with the basic dose constraints given in (2), additional binary variables are employed for modelling the dose-volume-tolerance relationships. To incorporate this concept into the model, let $\alpha_k,\ \beta_k \in (0, 1]$ for k in some index set K. (In our implementations, the cardinality of the index set K is typically between 3 and 10 but could be larger.) The following set of constraints ensures that at least $100\beta_k\%$ of the voxels in an organ-at-risk, OAR, receive dose less than or equal to $\alpha_k\, PrDose$. The symbols $y^{\alpha_k}_P$ and z^{OAR}_P denote binary variables.

$$\sum_{i\in\mathcal{B}}\sum_{j\in\mathcal{N}_i} D_{P,ij}w_{ij} \leq [\alpha_k\, PrDose]y_P^{\alpha_k} + \mathcal{D}_{max}z_P^{OAR},\ P\in OAR \tag{8}$$

$$\sum_{P\in OAR} y_P^{\alpha_k} \geq \beta_k|\, OAR\,| \tag{9}$$

$$y_P^{\alpha_k} + z_P^{OAR} = 1 \tag{10}$$

$$y_P^{\alpha_{k_1}} \leq y_P^{\alpha_{k_2}} \text{ for } \alpha_{k_1}\leq\alpha_{k_2} \tag{11}$$

Here, $\mathcal{D}_{max}$ is the maximum dose allowed for OAR (often determined by the maximum dose thought to be well-tolerated), and $\alpha_k,\ \beta_k$ combinations are patient and tumor specific. When the total dose received by a voxel P is less than $\alpha_k\, PrDose$, $y_P^{\alpha_k}=1$, and this contributes to a voxel count in Constraint (9). When it does not satisfy the dose bound, $y_P^{\alpha_k}=0$, and in this case the dose will be forced to be lower than the maximum dose tolerance allowed, $\mathcal{D}_{max}$, and $z_P^{OAR}=1$. Note that by using discrete variables to represent each voxel and controlling the number of points satisfying a certain dose level, we can impose strict dose-volume criteria within the solution space. This is in contrast to the common approach of incorporating "soft" dose-volume criteria into a composite objective function [113]. Langer[51] was the first to apply MIP ideas to model dose-volume relationships in conventional radiation therapy. For IMRT, the challenge is that the resulting problem instances are large-scale (involving hundreds of thousands or even millions of inequalities), and are computationally taxing and difficult to solve without the development of specialized algorithms [55].

Besides the commonly used least-squares dose deviation objective function, other objective functions have been used, including: minimizing the squared radiation dose to OARs, maximizing the minimum dose to tumor target, maximize/minimize weighted sum of doses to target and OARs. Other more complex biological objective functions — involving equivalent uniform dose (the p-norm or generalized mean value), tumor control probability, and normal tissue complication probability — have

also been proposed [64, 65, 77, 78, 79, 80, 81, 93, 94].

The MIP treatment planning models for real patient cases involve tens to hundreds of thousands of binary variables and constraints. Our experience is that the resulting MIP instances are intractable via commercial MIP solvers. However, we have observed that, by using specialized algorithms [55], clinically superior treatment plans can be obtained [56].

1.2.4 Computational Results for a Real Patient Case

We briefly describe a patient study. Input data includes 3D images of tissue to be treated. On these images, the planning target volume (PTV) is delineated, and contours of organs-at-risk (OAR) and normal tissue are outlined. In addition to these structures, a tissue ring of 5 mm thickness is drawn around the PTV. We call this ring the *critical-normal-tissue-ring*. In [31] it was demonstrated that this normal tissue construct can assist in obtaining conformal plans for radiosurgery. In [55, 56], we have shown its usefulness in designing conformal IMRT plans. For the results herein, depending on the volume of the anatomical structure, a $3-5$ mm voxel size (for dose computation) is used for setting up the MIP model instances.

For each beamlet, the dose per monitor unit intensity to a voxel is calculated. The total dose per unit intensity deposited to a voxel is equal to the sum of dose per intensity deposited from each beamlet. For the results described here, 16-24 coplanar fields of size $10\times10\ cm^2$ to $15\times15\ cm^2$ are generated as candidate fields, each of which consists of 400-900 $0.5\times0.5\ cm^2$ beamlets. This results in a large set of candidate beamlets used for instantiating the treatment planning models.

In [55], we study the effect of maximum beam angles allowed on plan quality. In [56], five objective functions are considered and contrasted on three different tumor sites to compare plan quality and to gain understanding of the steering effects of clinical objectives. Below, we illustrate the results for a head-and-neck case obtained

via multiple objectives.

Some common metrics for reporting quality of treatment plans include:

- Coverage — Coverage is computed as the ratio of the target volume enclosed by the prescription isodose surface to the total target volume. Coverage is always less than or equal to 1.
- Conformity — Conformity is a measure of how well the prescription isodose surface conforms to the target volume; it is computed as the ratio of the total volume enclosed by the prescription isodose surface to the target volume enclosed by this same surface. Conformity is always greater than or equal to 1.
- Homogeneity — The homogeneity index is defined as the ratio of the maximum dose to the minimum dose received by the tumor volume.
- Mean dose and maximum dose for each critical structure.
- Dose-volume histograms (volume receiving more than each given dose level) and isodose curves.

Observe that these metrics are not entirely independent. For example, while it is desirable to obtain a prescription isodose surface big enough to cover the target volume in order to ensure good coverage, it is also desirable to have this surface "small" in order to conform to the target volume. In addition, variations in conformity and coverage affect the amount of irradiation to nearby organs at risk, thus affecting dose distribution levels of these organs.

Head-and-neck tonsil cancer. We focus on a tonsil cancer case where the PTV is adjacent to the left submandibular salivary gland. The following structures with their respective clinical dose limits are considered. PTV should receive 68 Gy; left parotid: $30\% \leq 27$ Gy and $100\% \leq 68$ Gy; right parotid: $100\% \leq 15$ Gy; right submandibular

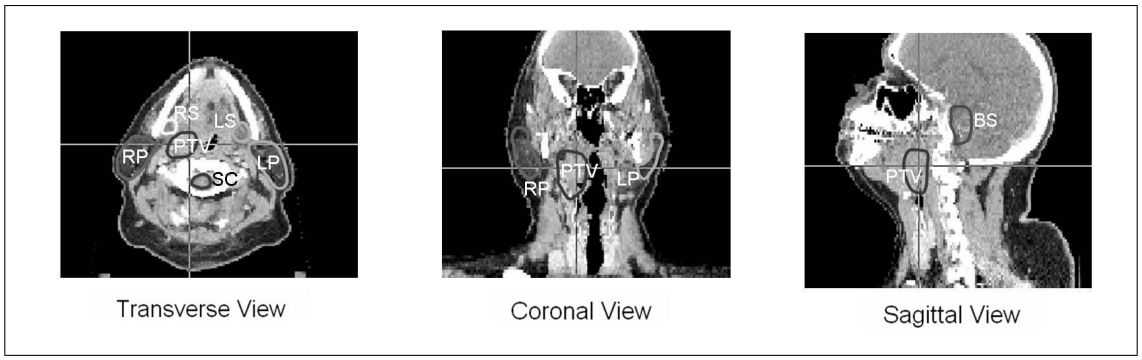


Figure 3: Anatomical structures for the head-and-neck. Notation: right parotid (RP), left parotid (LP), right submandibular gland (RS), left submandibular gland (LS), spinal cord (SP), brain stem (BS).

gland: $100\% \leq 30$ Gy; left submandibular gland: $10\% \leq 27$ Gy and $100\% \leq 68$ Gy; larynx: $80\% \leq 30$ Gy and $100\% \leq 55$ Gy; spinal cord and brainstem: $100\% \leq 45$ Gy.

A total of 1501 PTV voxels, 406 critical-normal-tissue-ring voxels, 3247 voxels for the OARs and 6416 normal tissue voxels were used to instantiate the MIP treatment model.

Here, we report the results for a plan with a maximum of 7 beams in which the objectives include minimizing the total dose to the critical structures and optimizing the PTV conformity. The results are based on the utilization of a specialized branch-and-bound MIP solver for large-scale external beam radiation [55] that is built on top of a general-purpose mixed integer research code (MIPSOL) [53]. Figure 4 shows the dose volume histograms, and Figure 5 shows the isodose curves. Compared to the clinical plan, we observe the following:

a. For all critical structures, the mean dose and max dose received are drastically less than the clinical plan.

b. For OARs that are close to the tumor volume, namely the left parotid ($< 10mm$) and the left submandibular gland ($< 10mm$), the mean dose received is significantly reduced (70% and 50%, respectively). The spinal cord enjoys moderate dose reduction (33%).

c. The coverage constraint and the objective helped in achieving 98% coverage. Underdose and overdose constraints kept minimum and maximum dose to the tumor relatively uniform, with a homogeneity index of 1.24. And the conformity objective helped to achieve a superior conformity value of 1.34. These all improve over the clinical plan, which had 97% coverage, and scores of 1.4 for homogeneity, and 1.6 for conformity.

d. The total overall monitor units of radiation from the MIP optimized plan is less than that from the clinical plan, indicating that the plan uses less radiation but yet can still deliver the required prescription dose to the tumor, thus sparing excessive radiation dose to the critical structures and normal tissue.

It is noteworthy that in contrast to the typically equispaced beams chosen when beam configurations are pre-selected, the optimal 7-beam plans obtained herein (that are considered clinically acceptable) do not have equispaced beams. Indeed, the optimal beam angles returned appear to be non-intuitive, and to depend on PTV size and geometry and the spatial relationship between the tumor and the critical structures.

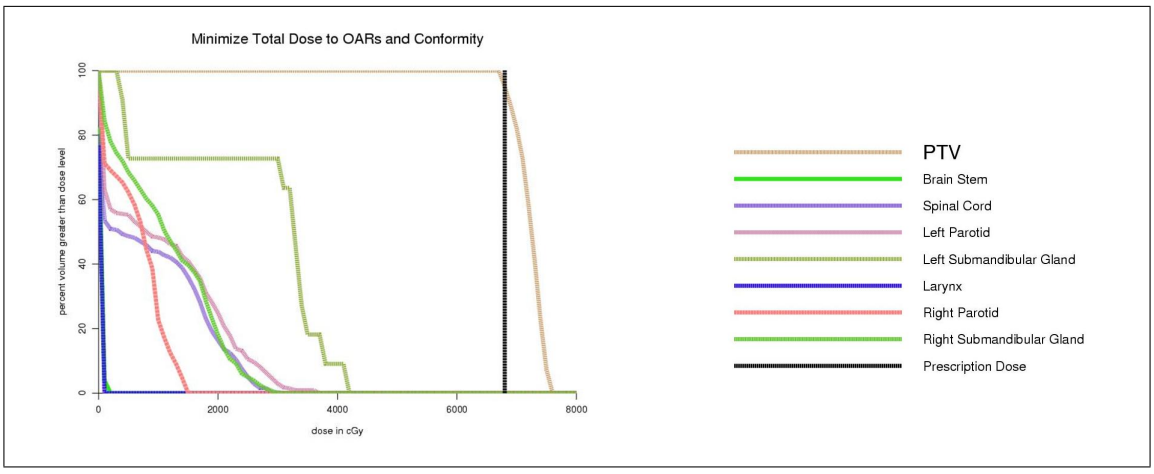


Figure 4: Dose-volume histogram for the head-and-neck for the MIP model with objective of minimizing the OARs dose and optimizing prescription dose conformity to tumor.

1.2.5 Other Mathematical Programming Approaches

As previously mentioned, linear and nonlinear programming have long been used for radiation therapy treatment optimization [5, 86, 87]. Lacking discrete variables, LP and NLP models typically use a pre-selected beam configuration, and focus on determining beam intensities. Below, we briefly outline some recent approaches in this area.

Simplified least-squares objective function and dose-volume constraints: In [86], using pre-selected beam angles, linear programming approaches were used to determine the associated optimal intensity map. The authors approximated the least-squares objective function measuring deviation of tumor voxel dose from prescribed dose via a piecewise linear function. They also utilized conditional value-at-risk (CVaR) constraints to control the mean dose received by subsets of voxels receiving the highest or lowest doses among all voxels in a given structure. Two forms of such constraints were used:

(i) lower α-CVaR: The average dose received by the subset of a target of relative volume 1-α receiving the lowest doses must be at least equal to L^{α}.

(ii) upper α-CVaR: The average dose received by the subset of a structure of relative volume 1-α receiving the highest doses may be no more than U^{α}.

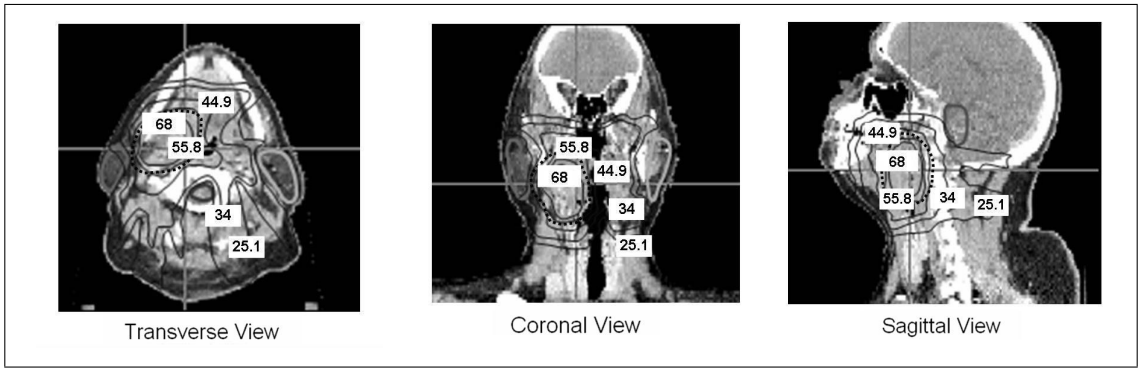


Figure 5: Isodose curves for the head-and-neck case. The critical-normal-tissue-ring is represented by the dotted curve.

CVaR constraints were originally proposed by Rockafellar and Uryasev [85] to formulate risk management constraints in terms of the tail means of distributions of financial risk. Mathematically, the upper α-CVaR constraint on a structure S is defined as

$$\bar{\zeta}_S^{\alpha}(w) + \frac{1}{(1-\alpha)|S|} \sum_{j \in S} \max\{0, D_P(w) - \bar{\zeta}_S^{\alpha}(w)\} \leq U_S^{\alpha}, \tag{12}$$

where U_S^{α} is an upper bound target, $D_P(w)$ is the total dose from intensity vector w for voxel P, and $\bar{\zeta}_S^{\alpha}(w)$ denotes the smallest dose level with the property that no more than $100(1-\alpha)$ percent of the structure S receives a larger dose. The authors showed that including such partial-volume constraints to bound the tail averages of the differential dose-volume histograms of structures helps to improve dose homogeneity to the target and to spare dose to critical structures.

Nonlinear programming approach: Sheperd et al [92] summarized several LP and NLP models for determining optimal intensity maps. To model dose-volume constraints, they applied a nonlinear error function approach. Their problem involved minimizing the standard objective of sum of the square differences between the prescribed and the actual doses over all of the voxels in the tumor, subject to two partial volume constraints — to OARs and to normal tissue. For an OAR S, the partial volume constraint defined on S was:

$$\sum_{P \in S} erf(D_P(w) - \Lambda_P) \leq \alpha |S| \tag{13}$$

where Λ denotes a selected dose limit and α denotes the fraction of the volume allowed to exceed this limit. The error function *erf(x)* realizing the partial volume constraints is a nonlinear function. (See fig. 4.4 in [92])

The authors also compared this with an MIP approach to model partial volume constraints on OARs, involving a simplified version of constraints (8) - (11) described above.

Direct aperture approaches via mixed integer programming: Preciado-Walters et al [83] formulated the treatment planning problem as a mixed integer program over a coupled pair of column generation processes: the first designed to produce intensity maps for the IMRT beamlet grid, followed by the second to specify protected area choices aiding in reducing the computational burden of enforcing the dose-volume restrictions on tissues.

Instead of determining the beamlet intensity for each beam, and then applying leaf-sequencing to determine delivery patterns, the planning involved first selecting a fixed set of deliverable beams. For each of these beams, the authors pre-determined heuristically a set of delivery patterns. They then introduced continuous nonnegative decision variables x_{jq} to represent the assigned intensity to whole pattern q of beam j. Then the dose at any voxel P, is calculated by

$$D_P = \sum_{j=1}^{n} \sum_{q \in \mathcal{Q}_j} a_{Pjq} x_{jq} \tag{14}$$

where $x_{jq} \geq 0$, $\mathcal{Q}_j$ is the set of patterns for beam j, and a_{Pjq} is the implied dose coefficient of pattern q from beam j at voxel P when the pattern q is constructed.

The resulting MIP model for treatment planning employed the objective function of maximizing the minimum tumor dose. Similar to the above MIP models, the constraints include upper and lower dose bounds on tumor voxels, and upper dose bounds for healthy tissues. Dose-volume constraints are formulated just as constraints (8) - (11) above.

1.3 Summary and Discussion

This article provides a brief overview of optimization issues in intensity-modulated radiation therapy, and summarizes our experience with an integer programming approach. The MIP model described allows simultaneous optimization over the space of beamlet intensity weights and beam angles. Based on experiments with clinical

data, this approach can return good plans that are clinically acceptable and practical. This work is distinguished from recent IMRT research in several ways. First, in previous methods beam angles are selected prior to intensity map optimization. Herein, we employ 0/1 variables to model the set of candidate beams, and thereby allow the optimization process itself to select optimal beams. Second, instead of incorporating dose-volume criteria within the objective function as in previous work, herein, a combination of discrete and continuous variables associated with each voxel provides a mechanism to strictly enforce dose-volume criteria within the constraints. The challenge of using MIP modelling for IMRT is that the resulting instances are very large-scale, and since general MIP is NP-hard, specialized algorithms designed to solve IMRT instances are required. Third, incorporating the critical-normal-tissue-ring can improve conformity in general tumor sites, without addition of other dose-shaping structures. In general, our MIP approach uses constraints to control a variety of clinical criteria (coverage, homogeneity, underdose to PTV, overdose to PTV, dose-volume limits on organs-at-risk and normal tissue), while assigning an objective to help with the solution search. The model can also be expanded to incorporate energy selection, couch angles and other treatment parameters.

Patient studies indicate that using the MIP approach, one can produce good clinical plans that aggressively lower OAR dose below pre-imposed levels without compromising local tumor control [56]. This is appealing since lower OAR dose should translate to lower normal tissue complication probability.

Computationally, the specialized optimization engine returns good feasible solutions within 30 minutes. We have performed standard leaf-sequencing techniques on the resulting optimal intensity map, and showed that returned plans are deliverable. The results provide evidence that the MIP approach is viable in producing good treatment plans that can potentially lead to significant improvement in local tumor control and reduction in normal tissue complication.

With pre-selected beam angles, other approaches such as linear programming [86] and nonlinear programming [92, 113] can be used for intensity map optimization. Comparisons are needed to gauge the quality of these plans versus those from MIP approaches. Direct aperture optimization [83] is appealing, since resulting segments are implementable directly. Again, comparisons are needed to determine the effectiveness, advantages and tradeoffs among different planning optimization methods.

Other computational challenges actively pursued by medical physics experts include image segmentation, planning under uncertainties, biological modelling, leaf-sequencing and treatment outcome analysis.

Acknowledgement

This research was partially supported by grants from the National Science Foundation, the National Institute of Health, and the Charles Edison Foundation.

CHAPTER II

QUADRATIC PROGRAMMING FOR FLUENCE MAP OPTIMIZATION IN INTENSITY MODULATED RADIATION THERAPY TREATMENT PLANNING

Eva K. Lee and Kyungduck Cha

Industrial and Systems Engineering

Georgia Institute of Technology

Atlanta GA, USA

evakylee@isye.gatech.edu

chacha2000@isye.gatech.edu

Abstract

In this chapter, we will describe the models and computational investigation related to fluence map optimization. In Section 2, the convex quadratic programming problem for FMO is introduced. The problem consists of a dose-based objective function with constraints for the upper and lower dose bound for each voxel, mean dose value received by a structure, and nonnegativity of beamlet intensities. In section 3, computational strategies including matrix reduction and post-processing scheme that can improve solution time are described. In section 4, the interior point method is described for solving the FMO problems. The computational results and plan quality for prostate and head-and-neck cases are summarized in section 5.

Our experiments show that the quadratic programming approach work successfully even when we model it with relatively simple parameters for the objective function. As observed from our computational experiments and the performance

of Cplex® barrier optimizer, a pure quadratic convex programming problem for FMO is unusually difficult due to its large-scale and the dense dose matrix. We demonstrate that a clever implementation of an interior point method can solve the quadratic convex programming problems very quickly. The key to this success involves a proper way of handling the large-scale and dense columns. Combining matrix reduction and post-processing scheme, we can solve these instances at faster rates still using our in-house QP-SOL solver.

2.1 Introduction

Intensity-Modulated Radiation Therapy(IMRT) is an advanced form of three-dimensional conformal radiation therapy. The treatment goal is to conform the spatial distribution of the prescribed dose to the planning target volume (PTV) while at the same time minimize the dose to the surrounding normal structures [39]. To determine a treatment plan, several optimization problems including beam angle selection, fluence map problem, and leaf-sequencing must be solved. In this chapter, we focus on the fluence map optimization (FMO) problem which is to find the optimal beamlet intensities given a set of pre-determined beam angles.

Depending on which mathematical model is employed for IMRT optimization problems, several solution methods have been used. These methods include linear programming [5, 87], mixed-integer programming [9, 10, 55, 56, 83], gradient method [12, 96, 113, 118], and simulated annealing [68, 75, 108]. Presently, the most common models in FMO problems are related to dose-based and dose-volume objective functions which are based on minimizing the weighted sum of the squares of the differences between the received dose and the prescribed dose or dose limit [14, 96, 113]. And these types of problems are solved by gradient methods, or via simulated annealing.

In gradient method, the optimization process performs iteratively by updating the search direction and step size of the next iteration point. There are several gradient algorithms classified by the way the search direction and the step size are determined. For radiation therapy planning, the Newton's method [12, 113] and conjugate gradient method [96] are commonly used. In addition to these algorithm, Zhang *et al* [118] implemented the steepest descent algorithm [84], and the scaled conjugate gradient algorithm [73] and tested their plan quality with lung and prostate cancer cases. Their experiments showed that Newton's method was by far the fastest, followed by the scaled conjugate gradient algorithm, while the steepest descent algorithm was the slowest.

Since the gradient methods are fast, they are often the favorite methods in IMRT optimization. However, when the gradient methods are used for FMO problems, multi local minima have been a concern [30]. To obtain an acceptable treatment plan, many studies have been focused on finding good objective function parameters, i.e., importance factors and the prescribed dose for input structures, by trial and errors or solving the problems iteratively by updating these parameters [112, 115, 116]. Even though some clinical studies [60, 114] showed that solutions from the gradients methods are sufficiently good, they did not prove the robustness of the gradient methods against multi local minima.

Since simulated annealing methods can find global or near-global minimum among multiple minima and do not require explicit forms of objective functions, these techniques have applied successfully to 3D-conformal radiation therapy (3D-CRT) and IMRT treatment planning. The basic aspect of simulated annealing is that it is analogous to thermo-dynamics, especially concerning the way that liquids freeze and crystallize, or that metals cool and anneal. Its solution time depends on a random generator that creates a new configuration (i.e. beam weights and beam orientations) and the temperature lowering schedule.

In early applications, classical simulated annealing provided by Kirkpatrick *et al* [49] was used for optimizing the beam weights and orientations in 3D-CRT treatment planning. Unfortunately, the method has been showed to be too slow in obtaining a solution due to the large number of iterations [107]. Thus, some specified knowledge to the desired solution was utilized to reduce the number of iterations [76, 108]. Mageras and Mohan [68] adopted the fast simulated annealing provided by Szu and Hartley [99], which uses the cauchy generator which is more efficient than the classical simulated annealing.

In general, inverse planning techniques such as the gradient methods are used for solving FMO problems; while fast simulated annealing techniques are employed to

find optimal combination of the pre-defined beam orientations [34, 98].

In this chapter, we use the interior point method to solve the *constrained* FMO problems. Since the work by Karmarkar [47], interior point methods have been successfully applied to solve large scale instances of linear programming problems and convex quadratic programming problems arising from industrial applications.

In Section 2, the convex quadratic programming problem for FMO is introduced. The problem consists of a dose-based objective function with constraints for upper and lower dose bound for each voxel, mean dose value received by a structure, and nonnegativity of beamlet intensities. In section 3, computational strategies including matrix reduction and post-processing scheme that can improve solution time are described. In section 4, the interior point method is described for solving the FMO problems. The computational results and plan quality for prostate and head-and-neck cases are summarized in section 5.

2.2 Modelling the Fluence Map Optimization

IMRT plan optimization begins with the patient information consisting of PTV, Organs-At-Risk (OARs), and normal tissues, which are delineated from CT or MRI images. In addition to these structures, we include the normal tissue ring drawn around the PTV with some thickness. Lee *et al* [54, 56] called this the *critical-normal-tissue-ring*, and demonstrated that this construction from normal tissues can assist in obtaining superior conformal plans for radiosurgery and IMRT plans. For simplicity, the critical-normal-tissue-ring is considered as one of the OARs in this study.

2.2.1 Input Data and Dose Calculation

To set up the model instance, each anatomical structure is discretized into *voxels*. We note that the size of the problem and the quality of the resulting plan will be affected by the choice of voxel size and the sampling scheme.

For FMO, the candidate beam angles are also predetermined. Then for each candidate beam, the shaped field of Multi-Leaf Collimator (MLC) is generally determined based on the PTV shape. In IMRT mechanism, the *modulated* beams are achieved by the MLC device attached to the head of the linear accelerator. 2D-shaped fields of MLC are represented by the discretized small grids, called *beamlets*. For each beamlet, the dose per monitor unit intensity to a voxel is calculated. We remark that computationally, dose calculation remains an important area of research. The total dose per intensity deposited to a voxel is given by the sum of dose per intensity deposited from each beamlet.

In our analysis, we employ the Computational Environment for Radiotherapy Research (CERR) software [31] to generate the patient information and dose data for treatment planning. CERR®, implemented by Deasy's group at Washington University in St. Louis, works on Matlab® environment and provides various functional components including the routines of importing and visualizing the AAPM/RTOG treatment planning format, the contouring tools to edit and/or create anatomical structures, the routines of generating the influence matrix with a beam setup, and the plan analysis tools to generate and visualize dose-volume histograms, and dose color-washes from the solution of treatment planning. In particular, the influence matrix of each patient was calculated with the Quadrant Infinite Beam (QIB) algorithm [46] which is a fast method of 3-D dose calculation based on Ahnesjö's approximation [3] of kernel dose deposition patterns.

Let B denotes the set of candidate beams and let N_i denotes the set of beamlets associated with beam $i \in B$. Given B, let the dose per intensity contribution to a voxel P from beamlet j in beam i be denoted as $D_{P,(ij)}$. Let $w_{ij}(>= 0)$ denotes the intensity of beamlet j in beam i. Then the total radiation dose at a voxel P, $D_P w$, is calculated as

$$D_P w = \sum_{i \in B} \sum_{j \in N_i} D_{P,(ij)} w_{ij}. \tag{15}$$

2.2.2 Quadratic Programming

2.2.2.1 Quadratic objective function models

The most common approaches in FMO problems have been related to dose-based and/or dose-volume objective functions. A typical dose-based objective can be defined as

$$\text{Min} \sum_{P \in PTV} \gamma_P (D_P w - d_P)^2 + \sum_{P \in OAR} \gamma_P H(w, d_P)(D_P w - d_P)^2 \tag{16}$$

where, d_P is the specified tolerance dose for a voxel P, γ_P is the weight parameter for a voxel P, and $H(w, d_P)$ is a step function defined as

$$H(w, d_P) = \begin{cases} 1, & \text{if } D_P w > d_P \\ 0, & O.W. \end{cases}$$

In case that $d_P = 0$ for $P \in OAR$, the step function, $H(\cdot, 0)$, is always 1 and this results in a simple quadratic equation, $\gamma_P (D_P w)^2$. For other cases where $d_P > 0$, we can also transform the step function into an explicit form. That is, by combining with some linear constraints, a dose-based objective (39) can be reformed to an explicit quadratic convex function as following:

$$\text{Min} \sum_{P \in PTV} \gamma_P (D_P w - d_P)^2 + \sum_{P \in OAR} \gamma_P (d_P v_P)^2$$

$$\begin{aligned} \text{S.T.} \qquad D_P w - d_P y_P &= d_P, \quad P \in OAR \\ v_P - y_P &\geq 0, \quad P \in OAR \end{aligned}$$

$$\begin{aligned} -1 \leq y_P &\leq M_P, \quad P \in OAR \\ 0 \leq v_P &\leq M_P, \quad P \in OAR \end{aligned}$$

where, y_P is a real-valued variable that measures the discrepancy between specified tolerance dose and actual dose, v_P is a real-valued variable that takes positive actual dose above the tolerance dose, and M_P is a sufficient large number to cover y_P and v_P for all $P \in OAR$.

Another popular quadratic objective function is a dose-volume based function. Its difference from the based-based objective function is on how the step function, H, is defined. That is, given $0 < d_P < d_{P_2}$,

$$H(w, d_P, d_{P_2}) = \begin{cases} 1, & \text{if } d_P < D_P w < d_{P_2} \\ 0, & O.W. \end{cases}$$

Since a step function, $H(w, d_P, d_{P_2})$, penalizes the voxels with the received dose values between d_P and d_{P_2}, the dose-volume based objective function appears to provide more flexibility. Indeed, proper parameters, d_P and d_{P_2}, which are often determined by lengthy trial-and-error approach, yield a better treatment plan. However, it does not guarantee that the objective functions are convex.

Let $f^{OAR:DV}(w) = \sum_{P \in OAR} \gamma_P H(w, d_P, d_{P_2})(D_P w - d_P)^2$. The following proposition shows that f is nonconvex.

Proposition 1: *$f^{OAR:DV}(w)$ is non-convex.*

Proof: This can be proved easily by selecting two points w_1 and w_2 as following:

$$w_1 \in \{w >= 0 : D_P w < d_P \ \forall \ P \in OAR\}$$

and

$$w_2 \in \{w >= 0 : D_P w > d_{P_2} \ \forall \ P \in OAR\}.$$

Note that $f^{OAR:DV}(w_1) = f^{OAR:DV}(w_2) = 0$ and there exists a voxel, P such that

$$H(\lambda w_1 + (1-\lambda)w_2, d_P, d_{P_2}) = 1 \text{ for } 0 < \lambda < 1.$$

Therefore, we have

$$f^{OAR:DV}(\lambda w_1 + (1-\lambda)w_2) \geq \lambda f^{OAR:DV}(w_1) + (1-\lambda) f^{OAR:DV}(w_2).\square$$

Consequently, when a gradient method is used to solve the problem of minimizing the dose-volume function, multi local minima traps have been observed. In our study, we focus on the dose-based objective function which can be transformed into an explicit quadratic convex program.

2.2.2.2 Constraints for FMO

In our FMO model, we will include some linear constraints used by Lee *et al* [55] and Romeijin *et al* [86]. Those constraints help not only to specify the clinical requirements, they will also compensate the weaknesses of the dose-based objective FMO model.

In many clinical practices, the minimum and maximum dose received by each voxel P need to be bounded. For a voxel P in PTV, lower and upper bounds, say L_P and U_P, are modelled as following:

$$L_P \leq D_P w \leq U_P. \tag{17}$$

And for a voxel $P \in OAR's$ and normal tissue, only upper bound constraints are added:

$$D_P w \leq U_P. \tag{18}$$

Clinically, upper and/or lower constraints are important as they prevent unexpected cold and hot spots in the resulting treatment plan. Moreover, proper dose limitation on each voxel guides the optimization solver to achieve the desired dose-volume restrictions. In terms of homogeneity index of PTV, defined as the ratio of the maximum dose to the minimum dose received by the tumor volume, can be determined by:

$$\text{homogeneity} \leq \max_{\{P \in PTV\}} \frac{U_P}{L_P}.$$

Further, in our model, we have also included the mean dose constraints for a critical structure, C_s, in PTV or OAR as follow:

$$\sum_{P \in C_s} D_P w \leq \zeta |C_s| \text{ or } \sum_{P \in C_s} D_P w \geq \zeta |C_s| \tag{19}$$

where ζ is a predetermined mean dose of C_s, and $|C_s|$ denotes the total number of voxels representing C_s. The mean dose constraints (19) can be extended to various forms of *partial mean dose constraints* based on the way the subset C_s is defined. For example, we can have

$$\sum_{P \in C_s^d} D_P w \leq \zeta^d |C_s^d|$$

where $C_s^d = \{P \in C_s : D_P w > d\}$ or $C_s^d = \{P \in C_s : D_P w < d\}$.

Table 1: Size and density of dose calculation matrix, D. Here, density of D is calculated by (# nonzero coeff.) / (size).

Cancer case	Size of D (# voxels) $\times$ (# beamlets)	# nonzero coeff. of D	Density of D (%)
Prostate	70,290 $\times$ 1222	8,632,385	10.05
Head-and-neck	114,000 $\times$ 1007	8,875,371	7.73

2.3 Matrix Reduction

The size of the dose matrix depends on the size of PTV and OAR, as well as the number of beam, thus beamlets, used. In general, the dose matrix of a real patient case is obtained by hundreds of thousand of voxels and thousands of beamlets. This results in a very large-scale optimization problem.

Moreover the primary and scatter dose from a beamlet to a voxel have been calculated for accurate dose calculation in order to reveal the actual dose distribution. The range of scatter dose may be estimated for the purpose of dose calculation. Even if we cut the pencil beams at a certain radial dose distance, say 1.5 cm [20], the dose matrix remains *relatively* dense. This is different from most of the large-scale instances arising from industries and logistics where they are mostly very sparse ($<$ 0.5 %.). Table 1 shows an example of the size and density of the dose matrices for two patient cases.

Since the scattered radiation dose can be described as an exponential function of radial distance between the central ray of the pencil beam and the voxel [3], it falls off radically as it moves away from the central ray. And for each beamlet, the magnitude of the scatter dose indicates the contribution relative to the primary. Therefore, a dose matrix typically consists of large numbers of very small coefficients while its maximum coefficient is often close to 1. Further, because of the noise occurred during computation, the maximum value may be comparatively greater than 1. Figure 6 shows the distribution of nonzero coefficients of a dose matrix obtained from a prostate cancer case. In this example, 48.67 % of the nonzero coefficients are less than 0.005

while 0.17 % are greater than 1.

We apply the matrix reduction scheme introduced by Lee *et al* [59] for brachytherapy into our IMRT FMO optimization. Specifically, we partition the dose matrix into submatrices based on the magnitude of the coefficients: for a chosen $\delta > 0$, the matrix D is split as $D = \overline{D} + \underline{D}$ where the elements of $\overline{D}$ and $\underline{D}$ make up of as following

$$\overline{d}_{ij} = \begin{cases} d_{ij}, & \text{if } d_{ij} \geq \delta \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad \underline{d}_{ij} = \begin{cases} d_{ij}, & \text{if } d_{ij} < \delta \\ 0, & \text{otherwise.} \end{cases}$$

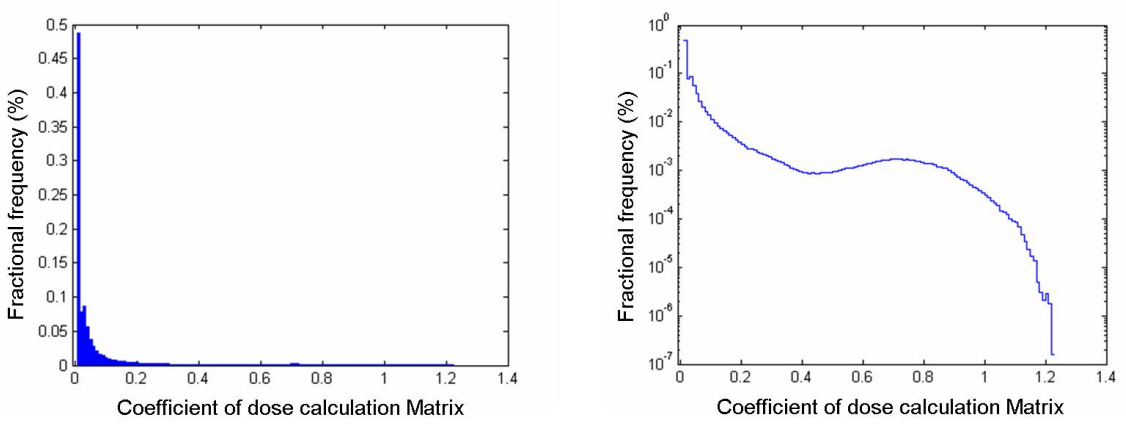


Figure 6: Histograms generated using the nonzero coefficients of a dose matrix for a prostate cancer case. In X-axis of both graphs, the size of disjoint category is 0.01. The right histogram with Y-axis scaled by *log* term is designed to display those extremely low levels of fractional frequencies.

Let P^s be the convex region satisfying the constraints $(17 - 19)$ defined at the section 2. Then the problems can be rewritten as

$$\text{Min}_{\{w \in P^s\}} \quad \gamma H(w,d)(Dw - d)^2 \tag{20}$$

$$\Updownarrow$$

$$\text{Min}_{\{w \in P^s\}} \quad \gamma H(w,d)\{(\overline{D}w - d)^2 + (\underline{D}w)^T(2(\overline{D}w - d) + \underline{D}w)\} \tag{21}$$

By ignoring submatix, $\underline{D}$, consisting of small coefficients, our reduction scheme is to solve the following *reduced problem*:

$$\text{Min}_{\{w \in \widetilde{P^s}\}} \ \gamma H(w,d)(\overline{D}w - d)^2 \tag{22}$$

Note that since the feasible region of constraints is also associated with the dose matrix, it should be changed into $\widetilde{P^s}$ in the reduced problem. In the case when δ is sufficiently small, say $\delta = 0.001$, we have observed that the second term of the system (21), $(\underline{D}w)^T(2(\overline{D}w - d) + \underline{D}w)$, does not affect the solution, thus an optimal solution to the reduced problem is almost identical to the solution of the original problem (20). Obviously, as δ increases, the solution to the reduced problem offers worse estimate to the original problem while it the submatrix $\overline{D}$ becomes sparser, which in terms, results in faster solution time. This tradeoff phenomenon will be illustrated in detail at section 2.5.

Let w^* and $\tilde{w}$ be optimal solutions corresponding to the original (20) and the reduced problem (22), respectively. In this work, we have also observed as following:

Observation 1: *If δ is relatively small, then $Dw^* \approx \overline{D}\tilde{w}$.*

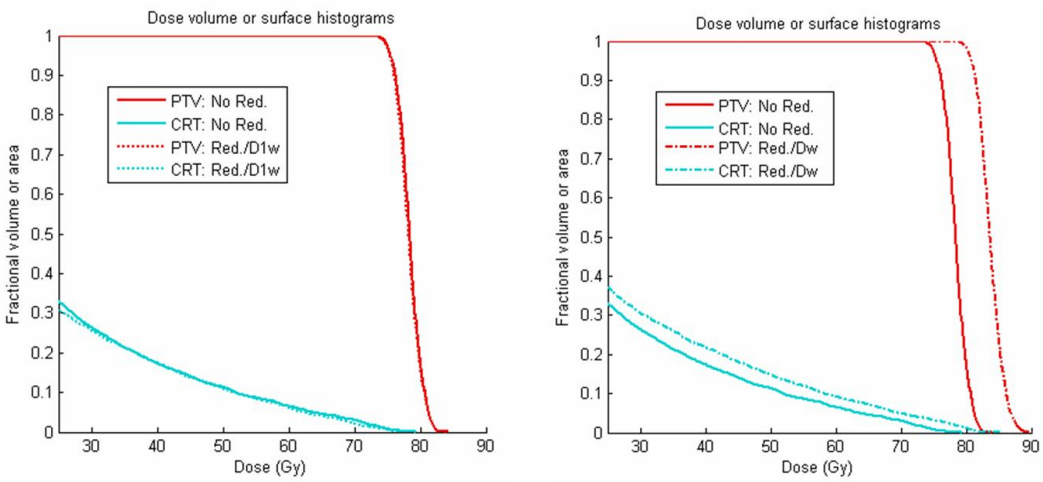
Therefore, if δ is relatively small, the DVH curves computed with $\overline{D}\tilde{w}$ should be expected to be almost identical to the DVH curves obtained from Dw^*. Figure 7 (a) shows the comparison DVHs computed with Dw^* and $\overline{D}\tilde{w}$ when $\delta = 0.04$. Moreover, since $\underline{D} \geq 0$ and $\tilde{w} \geq 0$, we have that

$$\overline{D}\tilde{w} \leq \overline{D}\tilde{w} + \underline{D}\tilde{w} = D\tilde{w}.$$

This implies that in case δ is relatively small, the relationship between Dw^* and $D\tilde{w}$ is

$$Dw^* \leq D\tilde{w}.$$

Figure 7 (b) illustrates DVH curves comparing between Dw^* and $D\tilde{w}$. In the graph, DVH curves generated with $D\tilde{w}$ appears to be shifted from curves with Dw^*.



(a) Dw^* vs $\overline{D}\tilde{w}$ (b) Dw^* vs $D\tilde{w}$

Figure 7: Dose volume histograms comparing solutions between w^* and $\tilde{w}$. DVH curves are plotted for the dose between 25 and 90 Gy. And $\tilde{w}$ is the solution to the reduced problem with $\delta = 0.04$.

Observation 2: *If δ is relatively small, then the shapes of DVH curves generated with $D\tilde{w}$ match closely to the ones generated with Dw^*.*

Our idea compensating the deviation from the matrix reduction is based on those observations.

Based on these observations, we will compensate the deviation from matrix reduction via a post-processing scheme. Specifically, we will do post-processing using the function of DVH values, $V_{(D,C,w)}(\alpha)$, defined as

$$V_{(D,C,w)}(\alpha) = \frac{|\{P \in C : D_P w \geq \alpha\}|}{|C|} \tag{23}$$

where C is a subset of PTV and/or OAR. The post-processing is to scale the solution of the reduced problem. For a given $[\beta_1, \beta_2]$, the scaling factor is computed as

$$\text{Min}_{\{\beta \in [\beta_1, \beta_2]\}} \left\{ \frac{V^{-1}_{(\overline{D},C,\tilde{w})}(\beta)}{V^{-1}_{(D,C,\tilde{w})}(\beta)} \right\} \tag{24}$$

where $V^{-1}_{(\cdot)}(\beta)$ is the inverse function of $V_{(\cdot)}(\alpha)$. Suppose that θ^* is the solution to the problem (24). Then the final scaled plan is $\theta^*\tilde{w}$.

2.4 Interior-Point Method for Quadratic FMO Problems

Our final model for FMO problems is a linear system with a quadratic convex objective function and a large number of linear constraints. Therefore, any gradient method, a favorite with clinicians, is not a good solution method anymore. In this chapter, we have applied an interior point method to solve these optimization problems. For simplicity, suppose that the QP problem is formulated as

$$\begin{aligned} \text{Min} \quad & x^T Q x/2 + c^T x \\ \text{S.T.} \quad & Ax = b \\ & 0 \le x \le u \end{aligned} \tag{25}$$

where x consists of (w, y, v), as introduced at Section 2. Clearly Q is a diagonal matrix. Thus we can readily extend the interior-point method for linear programming problems to these quadratic programming problems.

2.4.1 The Primal-Dual interior-point methods

Applying the logarithmic barrier function to eliminate the nonnegativity constraints, and after adding slack variables, s, to the upper bounds of x, the problem (25) is transformed into:

$$\begin{aligned} \text{Min} \quad & x^T Q x/2 + c^T x - \mu \textstyle\sum_{j=1}^{n} \ln x_j - \mu \sum_{j=1}^{n} \ln s_j \\ \text{S.T.} \quad & Ax = b \\ & x + s = u. \end{aligned} \tag{26}$$

Then the lagrangian function of transformed problem (26) is given by:

$$L(x,s,y,w,\mu) = x^T Qx/2 + c^T x - \varphi^T(Ax-b) - \lambda^T(u-x-s) - \mu\sum_{j=1}^{n}\ln x_j - \mu\sum_{j=1}^{n}\ln s_j. \tag{27}$$

The first-order necessary conditions for a stationary point of this lagrangian function (27) are

$$Ax = b \tag{28}$$

$$x + s = u \tag{29}$$

$$-Qx + A^T\varphi - \lambda + z = c \tag{30}$$

$$XZe = \mu e \tag{31}$$

$$S\Lambda e = \mu e. \tag{32}$$

where X, Z, S, and Λ are diagonal matrices with the elements x_j, z_j, s_j, and λ_j, respectively, and z is the vector of the dual slack variables.

In the application of interior point method to linear programming problems and convex quadratic programming problems, its key is to generate a sequence of feasible solutions $\{(x\ s\ \varphi\ z\ \lambda)^T\}$ to the conditions (28) - (32), converging to a solution with complementarity $x^T z + s^T\lambda <$ a predetermined tolerance. Since Karmarkar's projective algorithm [47], many algorithms have been proposed. In the view of the implementation for general-purpose linear programming problems, it is well-known that the commonest interior-point codes are based on the predictor-corrector algorithm [61, 62, 69]. (*In details, see the book written by Nocedal and Wright [82].*) Therefore, we extend and implement the predictor-corrector algorithm to solve the quadratic FMO problems.

2.4.2 Solving the Normal Equations

The biggest computational burden in interior-point methods is to solve the following normal equations:

$$\pi = (A\Theta A^T)^{-1}\xi \tag{33}$$

where, $\Theta = (Q + X^{-1}Z + S^{-1}\Lambda)^{-1}$. Clearly Θ is a diagonal matrix because Q is a diagonal in our problems. We assume that A has full row rank. This assumption makes $A\Theta A^T$ a positive definite symmetric matrix. Thus the equation (33) can be solved by using a cholesky factorization:

$$A\Theta A^T = LL^T$$

where, L is a lower triangular matrix. This approach has been applied successfully to large-scale problems by a *sparse column cholesky factorization* combined with a matrix ordering, e.g. *minimum local fill-in ordering* [63, 88].

However, in case that A contains any dense columns, the efficiency of cholesky factorization may be devastated by large numbers of fill-ins in $A\Theta A^T$ and L. Moreover, in the FMO problems, a dose matrix D contains thousands of relatively dense columns. This results in very significant fill-ins. The Schur Complement mechanism [4, 21] has generally been employed to handle dense columns in linear programming problems. In this study, we describe, based on the splitting scheme, how the Schur Complement can be applied to solve the quadratic programming problems in FMO.

Let $A = [A_s\ A_d]$, where A_s and A_d are matrices consisting of the spare and dense columns, respectively. Since Θ is diagonal, $A\Theta A^T$ can be rewritten as

$$A\Theta A^T = A_s\Theta_s A_s^T + A_d\Theta_d A_d^T. \tag{34}$$

Clearly, A_s also has full rank in our application to IMRT optimization in case that A has full row rank. Thus, by using sparse cholesky decomposition, the sparse part is computed as $A_s\Theta_s A_s^T = LL^T$. Applying Sherman-Morrison Formula to the inverse of equation (34), we have

$$(A\Theta A^T)^{-1} = (LL^T)^{-1} - (LL^T)^{-1}A_d(\Theta_d + A_d^T(LL^T)^{-1}A_d)^{-1}A_d^T(LL^T)^{-1}. \tag{35}$$

Finally, the normal equation (33) is represented as

$$(LL^T)\pi = \xi - A_d(\Theta_d + A_d^T(LL^T)^{-1}A_d)^{-1}A_d^T(LL^T)^{-1}\xi. \tag{36}$$

Therefore, the equation (33) can be solved by the Schur-complement method applied to following system:

$$\begin{bmatrix} A_s\Theta_s A_s^T & A_d \\ A_d^T & -\Theta_d \end{bmatrix} \begin{bmatrix} x \\ r \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}.$$

2.5 Computational Results and Clinical Comparisons

In our computational experiments, two patient cases of prostate and head-and-neck were selected, and the treatment input data were generated from CERR®. In addition to dose matrices, they include the geometry information of beams and the spatial and organ relational information. The patient information has been used to determine the model parameters for FMO problems.

2.5.1 Prostate cancer

The prostate cancer case contains of a primary PTV and a secondary PTV, labelled as PTV7560 and PTV4500, whose prescription doses are specified at 75.6 Gy and 45 Gy, respectively. As their coverage requirements, at least 95 % of each PTV should receive its prescription dose. To achieve the coverage, the model parameters of the objective function were manipulated. Consequently, the specified tolerance doses, d_P, at PTV are shifted by 3.5 % of the prescription dose and the weight parameters, γ_P, at PTV, are determined as

$$1/(\text{total number of voxels in a structure}).$$

Rectum and bladder are considered critical structures. It is specified that the dose to both critical structures cannot exceed 85 Gy and that 70 % of rectum and

bladder should receive less than 70 Gy and 80 Gy, respectively. Further, no voxel of the normal tissues can receive more than 95 Gy. For the model parameters of OAR, we use

$$d_P = 0 \text{ and } \gamma_P = 1/(\text{total number of voxels in a structure}).$$

Nine open 6 MV photon fields, selected by experienced clinicians, were used at the gantry angles of (0°, 40°, 80°, 120°, 160°, 200°, 240°, 280°, and 320°). Typically, each beam makes up of 20×20 beamlet grids with only beamlets delivering significant amount of dose to any PTV voxel included into the dose matrix. As a result, 1222 beamlets are selected for this patient. Recall that dose matrices associated with huge numbers of voxels are extremely large. Thus we have sampled voxels for each organ based on its importance. This results in reducing the size of dose matrix up to 25%.

We have generated the instance of a quadratic FMO problem with the dose matrix data and model parameters and solved the instance by our interior point codes, called *QP SOL*, as described in Section 2.4. Table 2 shows the size of the QP instance and its solution time running on a 850 MHz Pentium III Xeon machine. We also compare the solution time from our *QP SOL* code to the *barrier optimizer* from Cplex®. Note that our QP SOL code run much faster than Cplex® barrier algorithm on this instance.

Table 2: Size of instance and its solution time in prostate case

Cancer case	Size of instance (# rows) × (# columns)	Solution time (CPU Sec.) QP SOL / Cplex®
Prostate	35108 × 22712	397.32 / 11065.02

Using CERR®, we evaluated the solution plan via dose-washes, dose-volume histogram, and statistical indices. To validate the accuracy in modelling via sampling, these evaluations are computed with the original dose using the entire set of voxels (with no sampling). A dose-wash picture illustrates the overall dose distribution from the obtained optimal plan. As shown in Figure 8, the high doses indicated by reddish

color cluster on the primary tumor region of PTV7560, while the rectum and bladder are covered by low doses painted in shades of blue.

Figure 9 shows the dose-volume histograms for PTVs, critical structures, and all remaining normal tissues. The resulting plan is clinically acceptable and superior as it satisfies all the requirements in the planning directives. Specifically, 70% of rectum and bladder received less than 30 Gy, respectively.

Table 19 summarizes the dose statistics to the PTVs. Recall that the homogeneity is the ratio of maximum dose to minimum dose in the PTV. Conformity is a measure

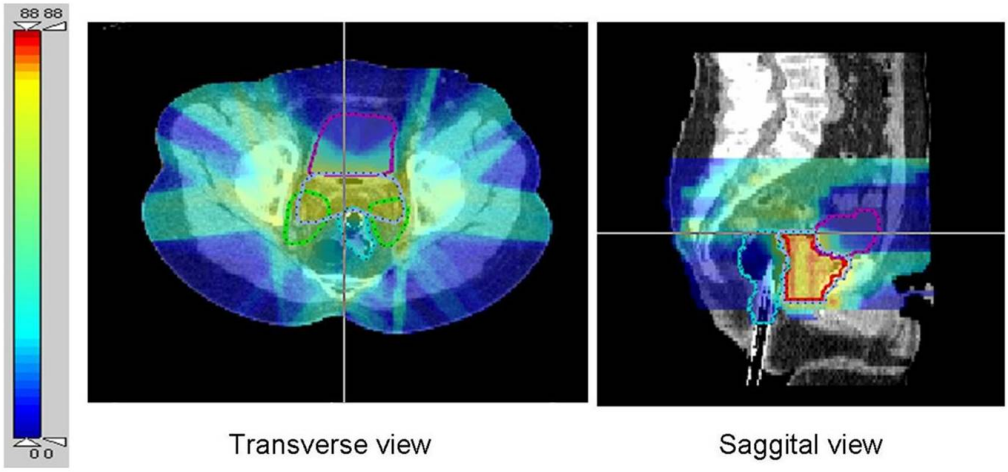


Figure 8: Dose-washes on a transverse and a sagittal view from the optimal plan for prostate patient. The color scale bar indicates the level of dose intensity.

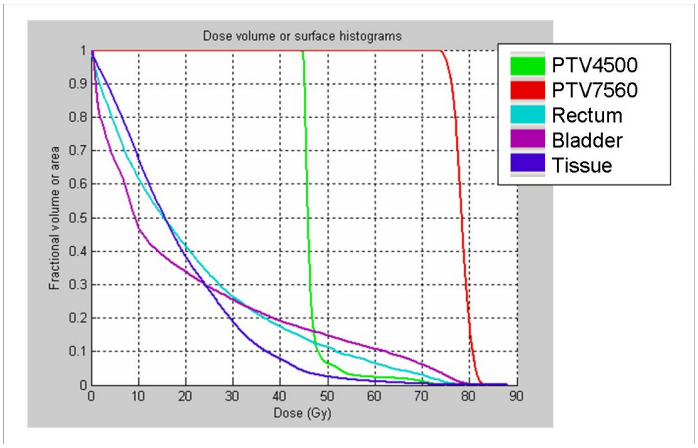


Figure 9: Dose-volume histogram generated from the optimal plan to prostate case

of how well the prescription isodose surface conforms to the target volume. That is, it is computed as the ratio of the total volume enclosed by the prescription isodose surface to the target volume enclosed by this same surface. Note that a superior plan results in homogeneity and conformity indices close to 1. Coverage is computed as the ratio of the target volume enclosed by the prescription isodose surface to the total target volume.

Table 3: DVH statistics of PTV7560 and PTV4500

	PTV7560	PTV4500
Max (Gy)	83.95	75.25
Mean (Gy)	78.465	46.862
Min (Gy)	71.15	44.25
Coverage	**0.95**	**0.972**
Homogeneity	**1.18**	
Conformity	**1.145**	

Table 4: DVH statistics of critical structures and unspecified tissues

	Rectum	Bladder	Tissue
Max (Gy)	79.35	82.05	88.05
Mean (Gy)	**21.348**	**20.159**	**18.478**
Min (Gy)	0.05	0.35	0.05
V_{20}	0.415	0.3395	0.3839
V_{40}	0.1749	0.1917	0.078
V_{60}	0.0649	0.1076	0.0108

Table 20 summarizes the dose statistics of rectum, bladder, and all remaining normal tissues. V_α is the DVH value of which a structure, C, receives greater than or equal to a dose level α for a solution plan, w: $V_\alpha = V_{(D,C,w)}(\alpha)$. Recall that $V_{(D,C,w)}(\alpha)$ is described at (23). As shown in Table 19 and 20, the optimal plan spares most of critical structures and normal tissues while meeting the clinical requirement for the PTVs.

For computational studies of the matrix reduction, we select several δ values such as 0.005, 0.01, 0.02, 0.03, 0.04, 0.05, and 0.06. That is, for each δ, the original dose

matrix D is reduced to the matrix $\overline{D}$ by taking the coefficients whose value are greater or equal to δ. Recall that 99.8 % of the coefficients of original dose matrix are less than 1, and these δ values are determined under the assumption that the maximum coefficient is 1. For each δ, we created the reduced quadratic problem using the same model parameters and solved it by using our QP SOL codes. Table 5 shows the density rates and the solution times for the original problem and the reduced problems. A density rate is defined as the density ratio of a reduced problem to the original problem. Figure 10 illustrates that the speed-up related to reduced rates appear to be superlinear. Here, a reduced rate is obtained by (1 - density rate) and the speed-up is calculated by

$$\frac{\text{(solution time of original problem)}}{\text{(solution time of a reduced problem)}}.$$

Let $\tilde{w}_\delta$ be the solution to equation (22) for each reduced problem using δ. Figure 11 shows the DVH graphs of each structure calculated from $D\tilde{w}_\delta$. All the DVH curves of the reduced problem from $\delta = 0.005$ are nearly identical to ones of the original problem. As δ value increases, the gaps of the DVH curves between the original and the corresponding reduced problem becomes bigger. However, the shape of all the

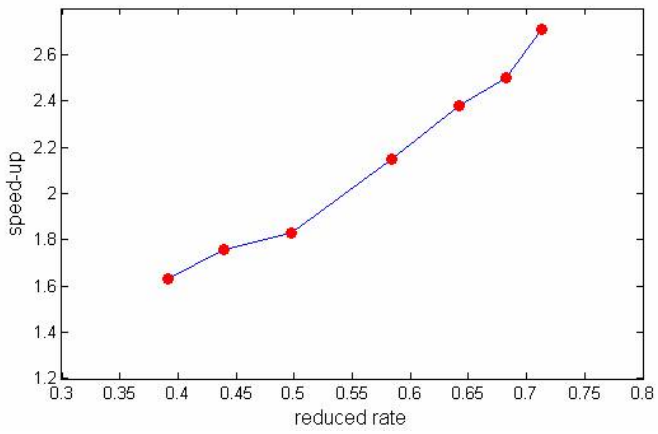


Figure 10: Relationship between solution time speed-up and the reduction of nonzero coefficients in the prostate case.

Table 5: Density rates and solution times for reduced problems in prostate case. The computations are run on a 850 MHz Pentium III Xeon machine. Here $\delta = 0$ corresponds to the original instance matrix

δ	Density rate	Solution time (CPU sec.)
0.0	1	397.3
0.005	0.61	243.2
0.01	0.56	225.4
0.02	0.50	217.0
0.03	0.42	184.4
0.04	0.36	167.0
0.05	0.32	158.7
0.06	0.29	146.8

DVH curves, except for $\delta = 0.06$, are marginally different from the original. Moreover for the DVH curves of PTV, the shapes are almost identical except a shift from the original. Based on these properties, the scaling factor can be determined by equation (24) to eliminate the effect of the matrix reduction. To compute equation (24), we set the parameters as follows:

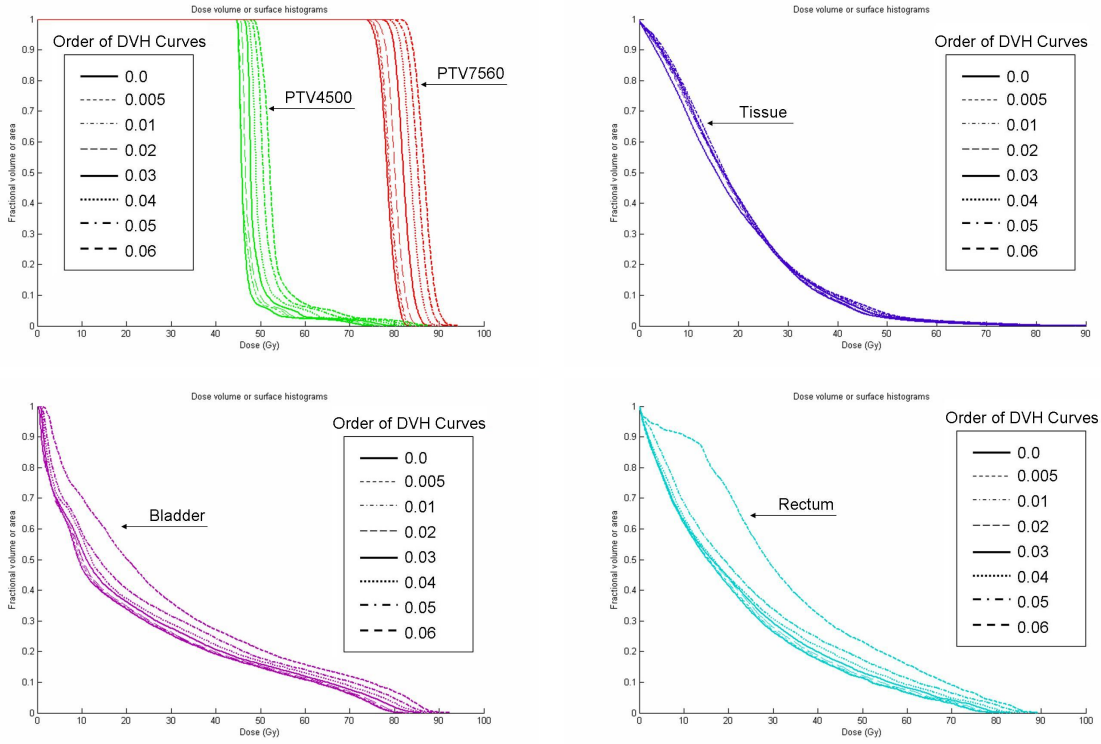


Figure 11: Comparison DVHs among δ values.

$$
\begin{aligned}
C &\leftarrow PTV7560 \\
\beta_1 &\leftarrow 0.93 \\
\beta_2 &\leftarrow 0.97.
\end{aligned}
$$

Since the equation (24) is extremely simple, we can obtain its solution, θ^*, in few seconds. Figure 12 shows the DVH graphs of bladder and rectum after post processing. And DVH graphs at Figure 13 compare the original problem with the reduced problems from each δ after post processing. That is, those DVH curves are calculated using $D\theta^*_\delta \tilde{w}_\delta$. Except from $\delta = 0.06$, the post-processed solutions are marginally different from the solution to the original problem. Fortunately, these differences happen around low level of doses. Combining the computational gain and the quality of the resulting plan, our matrix reduction scheme can improve the solution time by over 50%.

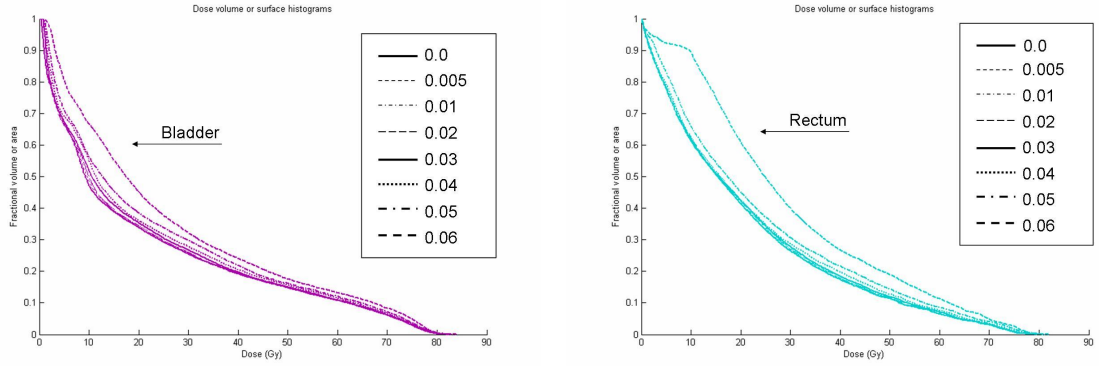


Figure 12: Comparison of DVHs of bladder and rectum after post processing

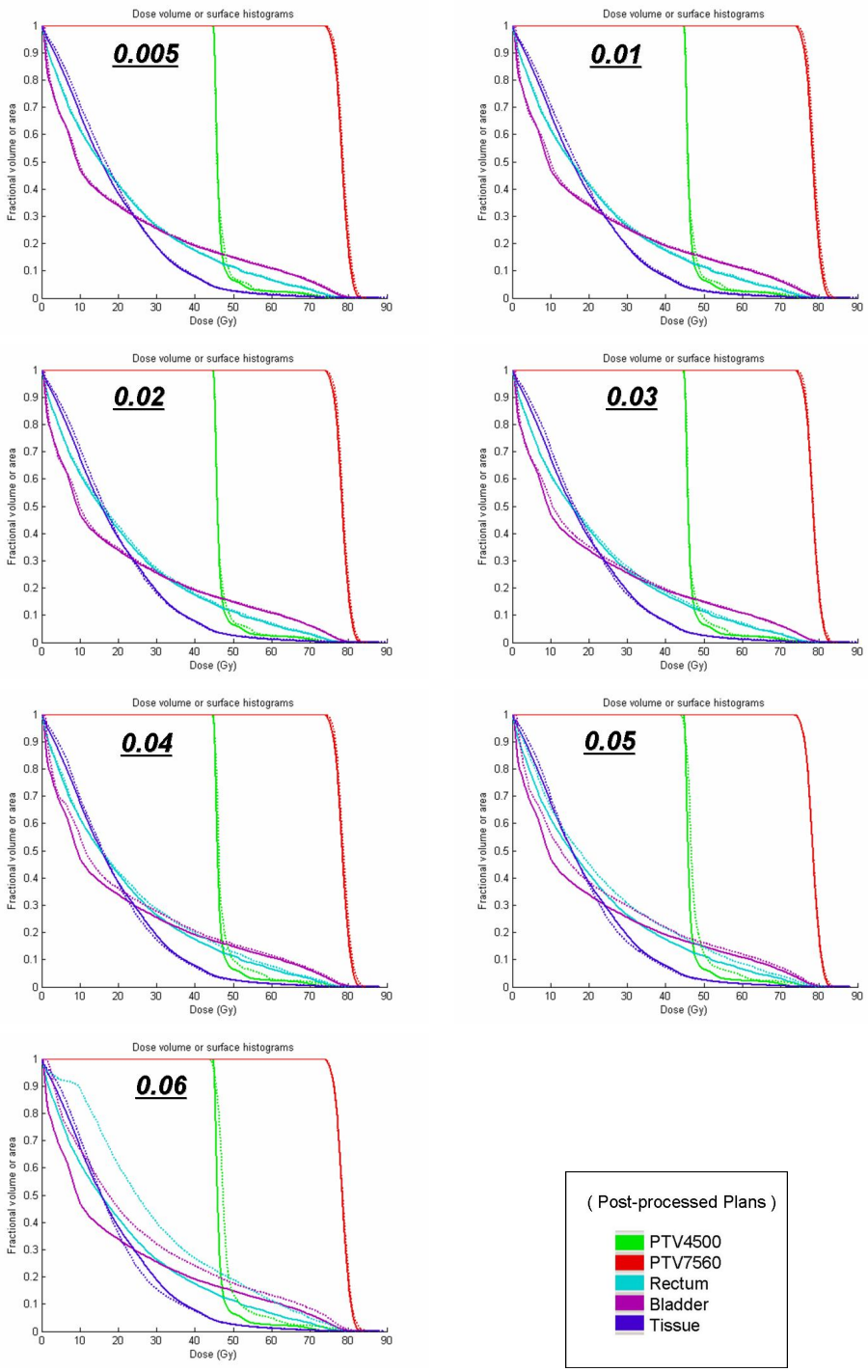


Figure 13: Comparison DVHs between the original problem and each reduced problem.

2.5.2 Head-and-Neck cancer

This patient case consists of a primary PTV and two secondary PTVs, whose prescription doses are specified at 72 Gy, 54 Gy and 49.5 Gy, with requirement that at least 95 % of each PTV should receive the prescription dose. Those PTVs are labelled as PTV7200, PTV5400, and PTV4950, respectively. PTV7200 and PTV5400 are wrapped inside PTV4950 to take into account of microscopic disease as well as the patient movement and setup uncertainty. The model parameters for PTVs are set to the same way as ones in the prostate case: the specified tolerance doses, d_P, are shifted by 3.5 % of the prescription dose and the weight parameter, γ_P, are determined to be 1/(total number of voxels in a structure).

The planning criteria also specifies that the dose to the spinal cord, brainstem, right parotid, and left parotid can not exceed 45, 54, 76.32, and 76.32 Gy, respectively, and that 50 % of left and right parotid should receive less than 30 Gy. Further, no voxel of the normal tissues can exceed 76.32 Gy. In the QP modelling, DHV restriction of each parotid, C_s, is replaced with a mean dose constraint:

$$\sum_{P \in C_s} D_P w \leq 30|C_s|.$$

For the model parameters of the OARs, any specified tolerance dose is set to 0. However, since this head-and-neck case is much convoluted in the view of the spatial geometry of OAR, the weight parameters are determined by combining the size of voxels and spatial information. All the voxels in OAR are first classified into several layer sets based on their distance from PTVs. Let l_P denotes an important factor of the layer assigned a voxel, P. In this work, we have designed 6 layers. Then the weight parameters of the OAR are computed as

$$\gamma_P = l_P/(\text{total number of voxels in a structure}).$$

Seven open 6 MV photon fields, selected by the clinicians, were used at the gantry angles (90°, 120°, 150°, 180°, 210°, 240°, and 270°). 1007 beamlets from these beams are included in the dose matrix. Similar to prostate case, the voxels are sampled based on their redundancy and importance. And we generated and solved the quadratic FMO problem on a 850 MHz Pentium III Xeon machine. As shown at Table 6, the instance from the head-and-neck case is much larger than the prostate case. This results in comparatively longer solution time by our QP SOL codes, whereas Cplex® cannot solve this instance and terminated as a result of *out of memory* occurring during the cholesky factorization.

Table 6: Size of instance and its solution time in the head-and-neck cancer case. On Cplex®, the problem is solved by *barrier optimizer*.

Cancer case	Size of instance (# rows) × (# columns)	Solution time (CPU Sec.) QP SOL / Cplex®
Head-and-neck	126,637 × 98,441	1604.4 / *Out of Memory*

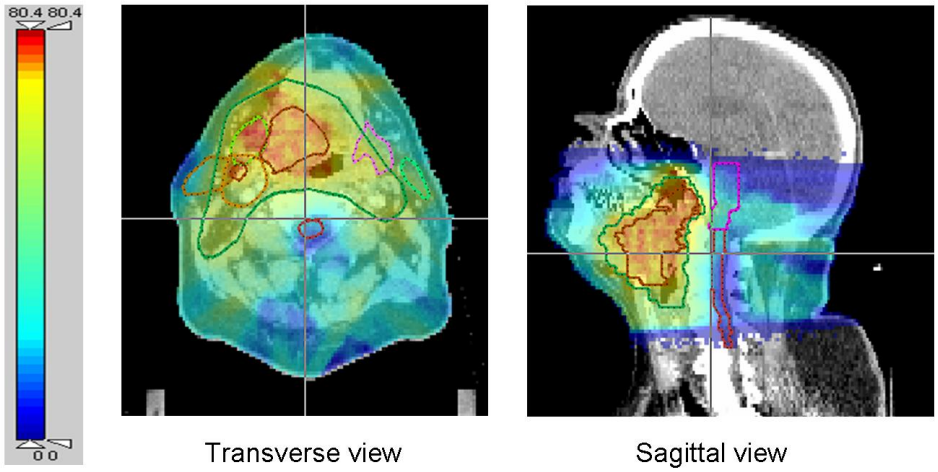


Figure 14: Dose-washes on a transverse and a sagittal view from the QP SOL optimal solution to the Head-and-Neck case. Color scale bar indicates the level of dose intensity.

The solution plan was evaluated by several clinical metrics, as discussed in the prostate case. These metrics are computed using the original dose data and the entire

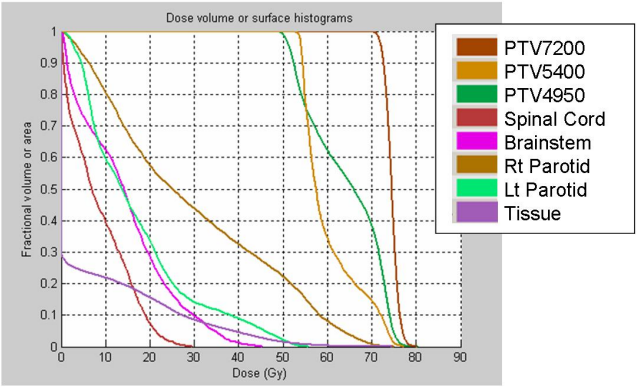


Figure 15: Dose-volume histogram generated from the optimal plan to the head-and-neck case

set of voxels (without sampling). The dose-washes at Figure 14 shows the overall dose distribution from the optimal plan. The sagittal view illustrates clearly that the plan delivers most of high doses to PTVs while sparing spinal cord, brainstem, and remaining tissues. Figure 15 shows the DVH graphs for PTVs, critical structures, and all remaining normal tissues. Table 7 and 8 summarize the dose statistics of the optimal solution to the original problem. These DVH graphs and statistics illustrate the solution to a quadratic FMO problem spares most of critical structures while satisfies all the requirements in the planning directives. In particular, mean dose constraints reflected well the DVH restrictions even though about half of the right parotid intersects with PTV 5400 and/or PTV4950. Note that for critical structures and normal tissues, the DVH curves and dose statistics are calculated using only the voxels that are not intersecting with the PTVs.

Table 7: DVH statistics of PTVs

	PTV 7200	PTV 5400	PTV 4950
Max (Gy)	80.25	77.15	80.45
Mean (Gy)	74.577	60.23	64.242
Min (Gy)	68.75	52.25	48.35
Coverage	**0.964**	**0.971**	**0.998**
Homogeneity	**1.167**		
Conformity	**1.444**		

Table 8: DVH statistics of critical structures and unspecified tissues

	Spinal cord	Brainstem	R. parotid	L. parotid	Tissue
Max (Gy)	**29.65**	**45.25**	**74.95**	**55.75**	**74.65**
Mean (Gy)	8.729	14.521	29.589	17	6.595
Min (Gy)	0.15	0.35	0.25	0.65	0.05
V_{20}	0.073	0.29	**0.575**	0.338	0.156
V_{30}	0	0.1	**0.44**	0.141	0.085
V_{40}	0	0.013	**0.327**	0.09	0.045
V_{50}	0	0	**0.224**	0.022	0.014

To study the effort of matrix reduction, we select several δ values such as 0.005, 0.01, 0.02, 0.03, 0.04, and 0.05. These δ values are determined under the assumption that the maximum coefficient of dose matrix is 1. Then a reduced quadratic problem for each δ is generated using the exact model parameters and solved by our QP SOL codes. Table 9 shows the density rates and the solution times for the original problem and the reduced problems. Figure 16 illustrates the speed-up with respect to the reduced rate. Since the PTVs are too large compared to the critical structures, the reduction rates are not as impressive as the prostate case. Consequently, the speed-up is not as prominent as in the prostate case.

Table 9: Density rates and solution times for reduced problems in head-and-neck case. The computations are run on a 850 MHz Pentium III Xeon machine. Here $\delta = 0$ corresponds to the original instance matrix

δ	Density rate	Solution time (CPU sec.)
0.0	1	1604.4
0.005	0.83	1421.1
0.01	0.80	1318.7
0.02	0.78	1328.6
0.03	0.68	1151.5
0.04	0.58	1060.7
0.05	0.52	1013.9

Upon evaluating the DVH graphs for each δ, we have observed the same behavior of DVH curves as in the prostate case. That is, as δ increases, the DVH curves are farther away from the original problem while their shapes remain comparable to the ones from the original problem. In this case, the threshold of DVH shape is 0.04.

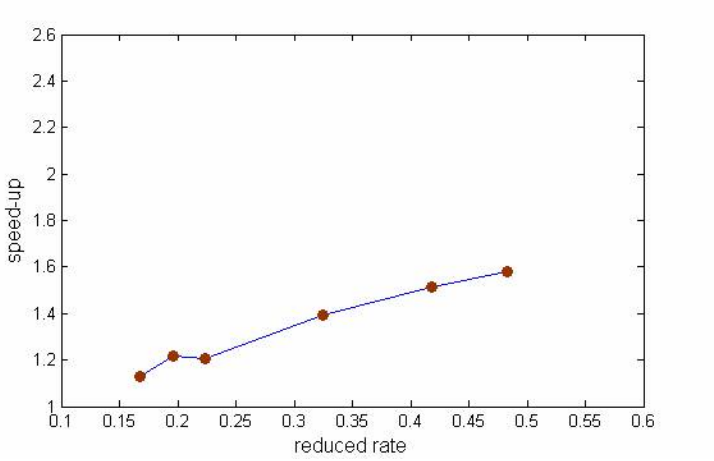


Figure 16: Relationship between solution speed-up and the reduced rate of nonzero coefficients in the head-and-neck case.

Based on these observations, the scaling factors are computed for equation (24):

$$\begin{aligned} C &\leftarrow PTV7200 \\ \beta_1 &\leftarrow 0.93 \\ \beta_2 &\leftarrow 0.97 \end{aligned}$$

DVH graphs in Figure 17 compare the original problem with the reduced problems for each δ after scaling. In the head-and-neck case, even though the post-processed solutions do not work as good as those from the prostate case, roughly speaking, the solutions, except for $\delta = 0.05$, are comparable to the original problem. Further, our matrix reduction scheme improves the solution by 37%.

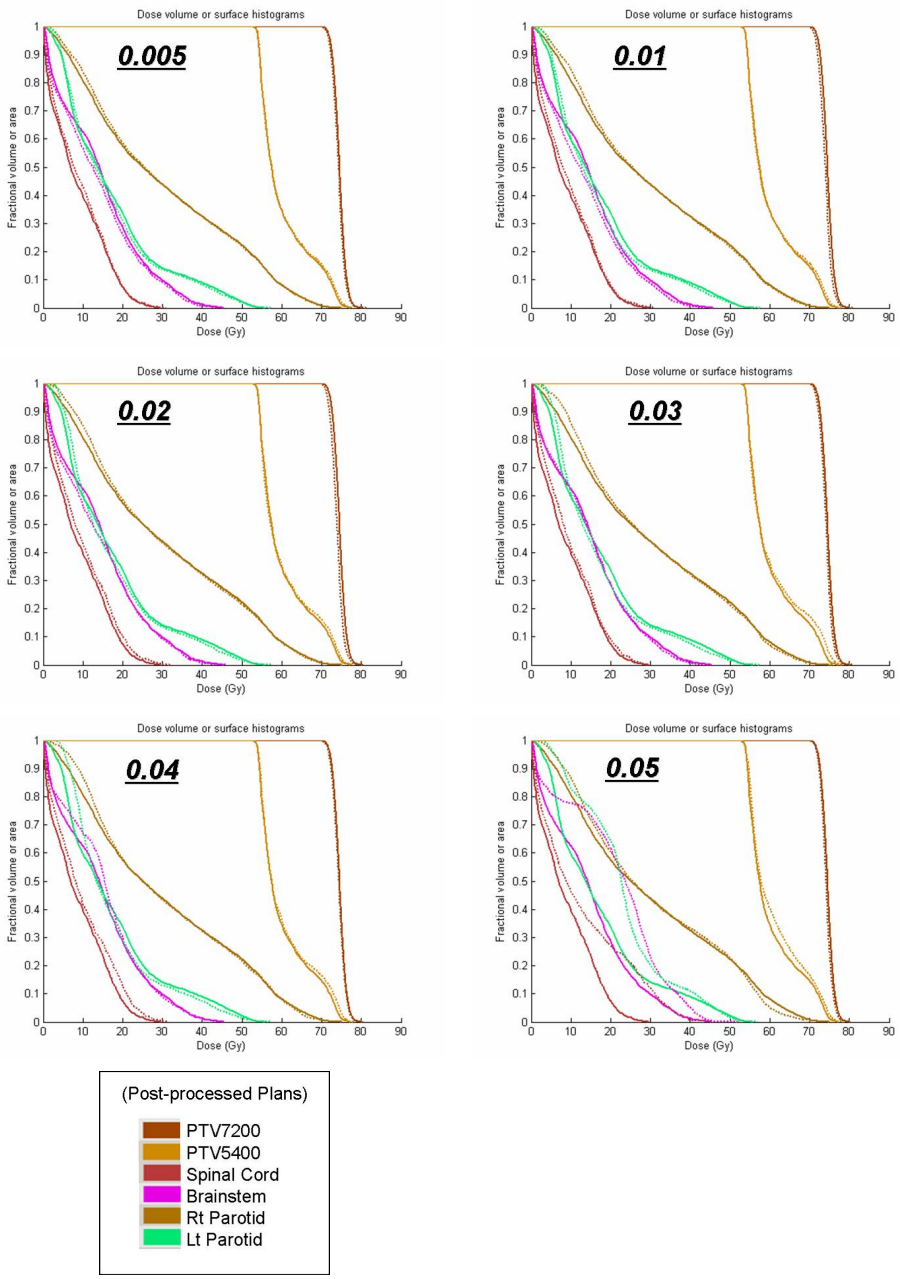


Figure 17: Comparison of DVHs between the original problem and the reduced problems.

2.6 Conclusion

In this Chapter, explicit quadratic convex programming problems combined with linear dose constraints with a dose-based objective function are formulated for IMRT fluence map optimization. Our experiments show the quadratic programming approaches work successfully even when we model it with relatively simple parameters for the objective function. Obviously, a better combination of model parameters can be found by manipulating the parameters iteratively which may lead to better treatment plans.

As observed from our computational experiments and the performance of Cplex® barrier optimizer, a pure quadratic convex programming problem for FMO is unusually difficult due to its large-scale and the dense dose matrix. We demonstrate that a clever implementation of an interior point method can solve the quadratic convex programming problems very quickly. The key to this success involves a proper way of handling the large-scale and dense columns. Combining matrix reduction and post-processing scheme, we can solve these instances at faster rates still using our in-house QP-SOL solver.

CHAPTER III

RESTRICTED COLUMN GENERATION APPLIED TO QUADRATIC MIXED INTEGER PROGRAMMING FOR INTENSITY MODULATED RADIATION THERAPY TREATMENT PLANNING

Kyungduck Cha and Eva K. Lee

Industrial and Systems Engineering

Georgia Institute of Technology

Atlanta GA, USA

chacha2000@isye.gatech.edu

evakylee@isye.gatech.edu

Abstract

Following our analysis on the fluence map optimization, in this chapter, we will focus on generation of beam angles and fluence map optimization simultaneously via a novel column generation scheme.

In Section 2 we describe briefly the input data and dose calculation for IMRT optimization problems and provide the mixed- integer quadratic model. Specially, the estimator of a DVH function is introduced to represent dose-volume constraints using linear inequalities. Solution strategy and computational advances are presented in Section 3 and 4. Specifically, we discuss some techniques that are useful in solving these large-scale intractable instances. In Section 5 we describe the computational experiences and present the clinical results illustrating the solution speed and resulting plan quality from our MIQP approach.

The proposed MIQP model allows one to optimize simultaneously the beam orientations over the entire space of the gantry angles and the beamlet intensity map. For computational efficiency, a restricted branch-and-price method is derived to solve these large-scale MIQP instances arising from clinical cases of prostate and head-and-neck cancer. Clinical experiments showed that the solutions from our MIQP are far superior than those plans obtained in by the current best-practice in the clinical environment.

Beyond drastically improving the computational time, the restricted branch-and-price method provides an additional bonus: every node consists of only a relatively small number of beam variables in the branch-and-bound framework. This implies that in the middle of the solution process, a feasible solution obtained can be clinically usable, and it can be readily obtained by a few branching steps of the beam variables that are in the active set

3.1 Introduction

Intensity-Modulated Radiation Therapy (IMRT) is an advanced technique of three Dimensional Conformal Radiation Therapy (3D-CRT). Its *modulated beam mechanism* may provide superior treatment plans to any conventional 3D-CRT, in terms of better conformal dose distribution of the prescribed dose to the Planning Target Volume (PTV) as well as much reduced dose to the surrounding normal structures. This sophisticated delivery mechanism requires solutions to several difficult optimization problems to be solved during the course of treatment planning.

Much research has focused on determining an optimal beamlet intensity map and/or optimal beam configuration (*or other beam delivery parameters such as field segments, coach angles, gantry angles, etc*). Since the attempt of Bahr et al [5] for radiotherapy planning in the late 60's, various mathematical models and solution approaches have been introduced to optimization problems arising in this area. These include linear (*or piece-wise linear*) programming [86, 87], gradient method [12, 96, 113, 118], simulated annealing [68, 75, 108], and mixed-integer programming (MIP) [9, 10, 83, 55, 57].

Presently, the most common models in the optimization of beamlet intensity map are related to dose-based and dose-volume objective functions which are based on minimizing the weighted sum of the squares of the differences between the received dose and the prescribed dose or dose limit [14, 39, 96, 113]. And gradient methods, known as efficient algorithms for unconstrained convex quadratic programming problems, are generally used to to solve these instances. While simulated annealing methods can find the global or near-global minimum among multiple minima and do not require the specific forms of objective functions, they are too slow to applied to a large-scale instances. Therefore, similar to 3D-CRT treatment planning, these techniques have popularly been used for the optimization for beam configuration problems.

MIP approaches, state-of-the-art computational techniques, allow one to consider

both beam angle and fluence map optimization simultaneously. For example, Lee *et al.* [55, 57] introduced this approach to determine the optimal beam angles and beam intensities. Preciado-Walters *et al.* [83] found a series of segments of multileaf collimators for radiation delivery.

In this chapter, we extend Lee's approach [55, 57] to a Mixed-Integer Quadratic Programming (MIQP) model where the objective function takes on a dose-based objective function. The computational approach focuses on the design of an efficient a branch-and-price approach to solve these MIQP instances. Branch-and-price has been considered an efficient solution approach to solve large-scale MIP instances. Since Gilmore and Gomory [37, 38]'s researches on the cutting stock problem, branch-and-price method has been successfully applied to various classes of MIP instances including vehicle routing problem [32, 33], aircraft fleeting and routing problem [6], single machine problem [101], bin packing and cutting stock problems [102, 103], minimum cut clustering problem [45], and generalized assignment problem [89], etc. In our context, we develop a branch-and-price algorithm for solving large-scale MIQP arising from IMRT.

In Section 2 we describe briefly the input data and dose calculation for IMRT optimization problems and provide the mixed- integer quadratic model. Specially, the estimator of a DVH function is introduced to represent dose-volume constraints using linear inequalities. Solution strategy and computational advances are presented in Section 3 and 4. Specifically, we discuss some techniques that are useful in solving these large-scale intractable instances. In Section 5 we describe the computational experiences and present the clinical results illustrating the solution speed and resulting plan quality from our MIQP approach.

3.2 Modelling IMRT Problem

IMRT plan optimization begins with the patient information consisting of PTV, Organs-At-Risk (OARs), and normal tissues, which are delineated from CT or MRI images. In addition to these structures, we include the normal tissue ring drawn around the PTV with some thickness. Lee *et al* [54, 56] called this the *critical-normal-tissue-ring*, and demonstrated that this construction from normal tissues can assist in obtaining superior conformal plans for radiosurgery and IMRT plans. For simplicity, the critical-normal-tissue-ring is considered as one of the OARs in this study.

3.2.1 Input Data and Dose Calculation

To set up the model instance, each anatomical structure is discretized into *voxels*. We note that the size of the problem and the quality of the resulting plan will be affected by the choice of voxel size and the sampling scheme.

For simultaneously beam angle and fluence map optimization, a large set of candidate beam angles are predetermined (30-40). Then for each candidate beam, the shaped field of Multi-Leaf Collimator (MLC) is generally determined based on the PTV shape. In IMRT mechanism, the *modulated* beams are achieved by the MLC device attached to the head of the linear accelerator. 2D-shaped fields of MLC are represented by the discretized small grids (called beamlets). For each beamlet, the dose per monitor unit intensity to a voxel is calculated. Note that in terms of dose computation, efficient dose calculation approach is an important research area because dose computations (i.e., Monte Carlo method) associated with a large number of beamlets are computationally very expensive. The total dose per intensity deposited to a voxel is equal to the sum of dose per intensity deposited from each beamlet.

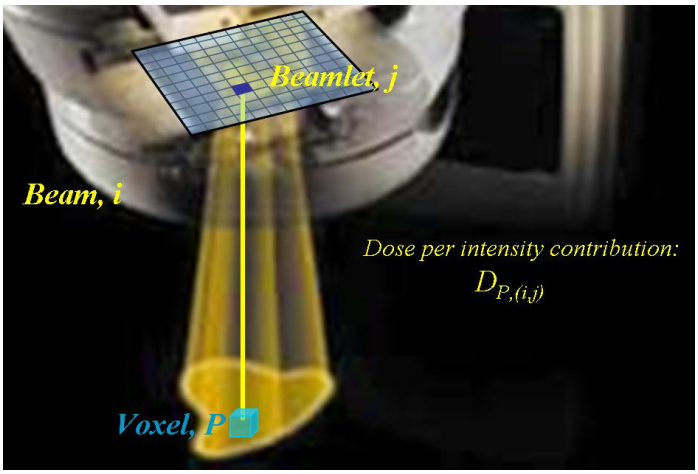


Figure 18: Dose per intensity contribution to a voxel P from beamlet j in beam i.

3.2.2 Basic Quadratic Programming

Let B denotes the index set of predetermined candidate beams and let N_i denotes the index set of beamlets associated with beam $i \in B$. Given B, let the dose per intensity contribution to a voxel P from beamlet j in beam i be denoted as $D_{P,(ij)}$. Figure 18 illustrates the scheme of this dose calculation. Let w_{ij} denotes the intensity of beamlet j in beam i. Let $w_{ij}(>= 0)$ denotes the intensity of beamlet j in beam i. Then the total radiation dose at a voxel P, $D_P w$, is calculated as

$$D_P w = \sum_{i \in B} \sum_{j \in N_i} D_{P,(ij)} w_{ij}. \tag{37}$$

Clinically, the minimum and maximum dose, say L_P and U_P, received by a voxel P can be modelled as:

$$L_P \leq D_P w \leq U_P. \tag{38}$$

It is ideal that for $P \in OAR$ or normal tissue, $L_P = 0$. For example of a head-and-neck cancer case, every voxel in the *spinal cord* should receive less than 45 Gy. Here, Gy stands for *Gray* which is the energy (Joules) deposited locally per unit mass (Kg). Besides their clinical requirements, proper upper and lower bounds prevent

unexpected hot and cold spots, respectively. And in terms of homogeneity index which is defined as the ratio of the maximum dose to the minimum dose received by the tumor volume, the bounds, L_P and U_P, for PTV constrain the upper bound of homogeneity in a solution plan. That is,

$$\text{homogeneity} \leq \max_{\{P \in PTV\}} \frac{U_P}{L_P}.$$

The dose-based objective function for IMRT treatment planning is of the form

$$f^{obj} = \sum_{P \in PTV} \gamma_P (D_P w - d_P)^2 + \sum_{P \in OAR} \gamma_P H(D_P w - d_P)(D_P w - d_P)^2 \tag{39}$$

where, d_P is the specified tolerance dose for a voxel P, γ_P is the weight parameter for a voxel P, and $H(D_P - d_P)$ is a step function defined as

$$H(D_P w - d_P) = \begin{cases} 1, & \text{if } D_P > d_P \\ 0, & \text{Otherwise.} \end{cases}$$

In case that $d_P = 0$ for a $P \in OAR$, its step function H_P can be ignored and this results in a simple quadratic term, $\gamma_P (D_P w)^2$. For the other cases where $d_P > 0$, we can consider the relationship between volume of the structure and dose received, and transform its step function H_P explicitly. Consequently, the objective function (39) can be transformed into an explicit quadratic convex programming with some constraints as following:

$$\text{Min} \sum_{P \in PTV} \gamma_P (D_P w - d_P)^2 + \sum_{P \in OAR} \gamma_P (d_P v_P)^2 \tag{40}$$

S.T.

$$D_P w - d_P y_P = d_P, \quad P \in OAR \tag{41}$$

$$v_P - y_P \geq 0, \quad P \in OAR \tag{42}$$

$$-1 \leq y_P \leq M_P, \quad P \in OAR \tag{43}$$

$$0 \leq v_P \leq M_P, \quad P \in OAR \tag{44}$$

where, y_P is a real-valued variable that measures the discrepancy between specified tolerance dose and actual dose, v_P is a real-valued variable that takes positive actual dose above the tolerance dose, and M_P is a sufficient large number to cover y_P and v_P for $P \in OAR$.

3.2.3 Dose-Volume Restrictions

In radiation therapy, it has believed that radiation goes straight through human body and it is impossible for it to reach PTV without harming OAR and normal tissue surrounding PTV. Consequently, a part of OAR and/or normal tissue may be sacrificed in some levels while covering PTV by desired level of prescription dose.

In clinical practice, thus, it is common for oncologists to specify the percentage of volume a critical organ is allowed to receive a certain among of radiation dose. This specification is known as the Dose-Volume (DV) restriction. That is, a DV restriction is typically stated as at most $\beta \times 100$ % of a critical organ is allowed to receive the dose greater than or equal to α Gy. For example, in a head-and-neck cancer case, at most 50 % of the *parotid* can receive a dose greater than or equal to 30 Gy. For PTV, DV restrictions are sometimes be employed to the upper dose tale (higher than the prescription doses). Moreover, several levels of DV constraints can be specified on a single structure to achieve better quality of treatment plan.

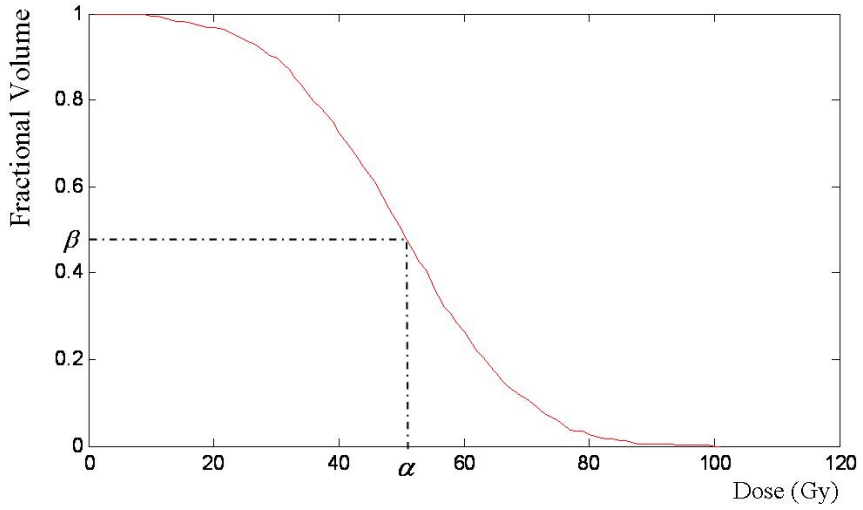


Figure 19: DVH graph and DV constraint represented by $V_\alpha(\ \cdot\ ,\ w)$ at a specified dose level α.

Clinicians use the Dose-Volume Histogram (DVH) to evaluate the quality of a treatment plan. Let $V_\alpha(S_i, w)$ be the value of DVH in which a critical organ, $S_i \subseteq OAR$, receives greater than or equal to a dose level α from a beamlet intensity map, w. Then, $V_\alpha(S_i, w)$ can be calculated as

$$V_\alpha(S_i, w) = \frac{|\{P \in S_i : D_P w \geq \alpha\}|}{|S_i|} \tag{45}$$

where, $|\ \cdot\ |$ denotes the cardinality. As shown in Figure 19, given $D_P w,\ \forall\ P \in S_i$, a typical DVH graph is generated by plotting $V_{\alpha_k}(S_i,\ w)$ sequentially at discretized dose level α_k. And at a specified dose level α, DV constraint can be represented by limiting $V_\alpha(S_i, w)$:

$$V_\alpha(S_i, w) \leq \beta. \tag{46}$$

Some investigators [51, 55, 57] suggested mixed-integer programming (MIP) to formulate DV constraints explicitly. Note that their explicit formulations consists of binary (or integer) variables representing $|\{P \in S_i : D_P w \geq \alpha\}|$. Because of well-known difficulty in solving MIP problems, specialized techniques are required to solve

these instances to produce plans with superior clinical quality.

On the other hand, medical physicists have considered the DV constraints implicitly by using the dose-volume objective functions [113] to penalize only those voxels that receive the doses between two predetermined values d_1 and d_2 and solving them via gradient methods, known to be an efficient algorithm for solving unconstrained convex quadratic programming problems. However, since there exist multiple local extrema on these functions [22, 30, 114], the final solution may not be a global extrema.

In this chapter, we introduce an estimator of DVH value derived from partial dose in a dose interval: instead of counting the number of voxels in the set, $\{P \in S_i : D_P w \geq \alpha\}$, at a dose level α, a DVH estimator is computed based on the relationship between two partial mean doses in interval $(\alpha - \delta_1,\ \alpha + \delta_2)$. Note that this is the extension of single mean tail dose [86], in which DV constraints can be approximated with linear convex constraints, free from integer variables.

Let $\tilde{V}_{(\alpha,\delta)}$ denotes the DVH estimator in dose interval $(\alpha - \delta_1,\ \alpha + \delta_2)$ at α. Then for given δ_1 and δ_2, $\tilde{V}_{(\alpha,\delta)}(S_i, w)$ is defined as

$$\tilde{V}_{(\alpha,\delta)}(S_i, w) = \frac{\sum_P \max\{0, (D_p w - (\alpha - \delta_1)\} - \sum_P \max\{0, (D_p w - (\alpha + \delta_2)\}}{(\delta_1 + \delta_2)|S_i|}. \quad (47)$$

As shown in Figure 20, DVH curves are generated by plotting $\tilde{V}_{(\alpha,\delta)}(S_i, w)$ at each discretized dose level α. Figure 21 plots the comparison of DVH curves generated by the conventional DVH values and the estimators of DVH value with $\delta_1 = \delta_2 = 0.5$ respectively. Both curves are almost identical to each other. Clearly, in the case that δ_1 and δ_2 are sufficiently small and $|S_i|$ is sufficiently large, we have

$$\tilde{V}_{(\alpha,\delta)}(S_i, w) \cong V_\alpha(S_i, w).$$

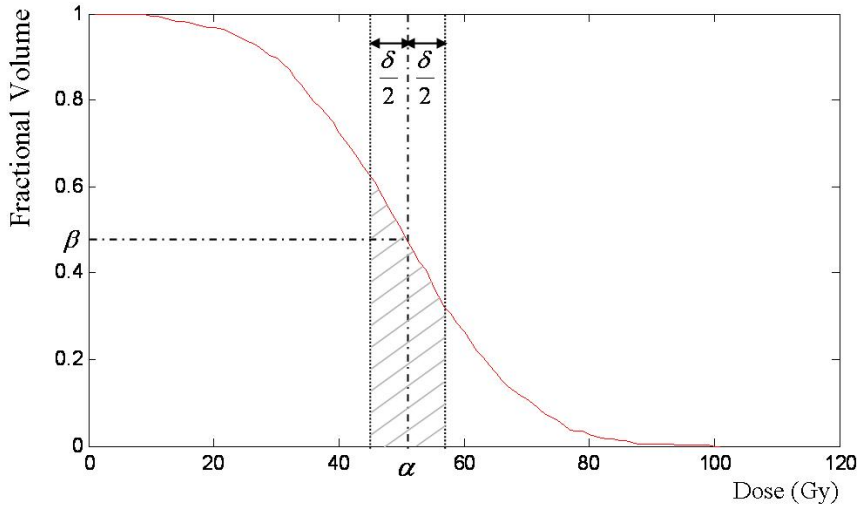


Figure 20: DVH graph generated by estimators in dose interval $(\alpha - \delta/2,\ \alpha + \delta/2)$ at α. Here, δ_1 and δ_2 are set to $\delta/2$ for this example. Note that the area covered by oblique lines is computed by $\sum_P \max\{0, (D_p w - (\alpha - \delta/2)\}/|S_i| - \sum_P \max\{0, (D_p w - (\alpha + \delta/2)\}/|S_i|$.

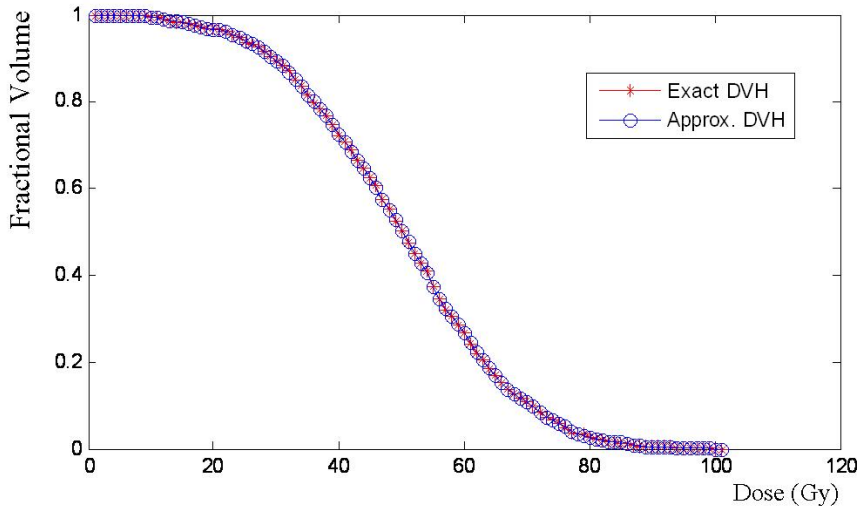


Figure 21: Comparison of DVH graph between the conventional calculation and the estimator in dose interval. Here, $\delta_1 = \delta_2 = 0.5$.

Consequently, the inequality constraint (46) can be approximated by two partial mean doses in the interval $(\alpha - \delta_1,\ \alpha + \delta_2)$ as following:

$$\sum_{P} \max\{0, (D_p w - (\alpha + \delta_2)\} \leq \xi \quad (48)$$

$$\sum_{P} \max\{0, (D_p w - (\alpha - \delta_1)\} \leq \xi + \beta\{(\delta_1 + \delta_2)|S_i|\}. \quad (49)$$

3.2.4 Beam Angle Selection

It is generally believed that increasing the number of beam angles would improve the quality of treatment plan. However, this may increase beam delivery time which results in more cost of the treatment and more chances for potential errors connected with patient setup, organs motions, patient movement and increased discomfort. Moreover, some researchers [74, 106] pointed out that longer delivery time may affect cell survival and tumor control probability. Thus, in the clinical and practical setting, the number of beams used in a final plan needs to be limited by a relatively small number. For example, at most 7 beams are allowed to be used in most head-and-neck cases.

Let x_i be a binary variable denoting the usage of beam i in a treatment plan. That is, if beam i is to be used then $x_i = 1$, otherwise $x_i = 0$. Given B_{max}, the maximum number of beams desired in an optimal plan, the constraint of the beam selection is

$$\sum_{i \in B} x_i \leq B_{max}. \quad (50)$$

Recall that B is the index set of candidate beams. For the optimization problem of beam orientation, the candidate beams are generally predetermined by discretizing the entire space of the beam orientation in increments of a spacing unit, θ (for example, 5^o or 10^o). And when beam i is not used, the beamlets corresponding to this beam should be zero. The following constraints represent the relationship between a beam and its associate set of beamlets:

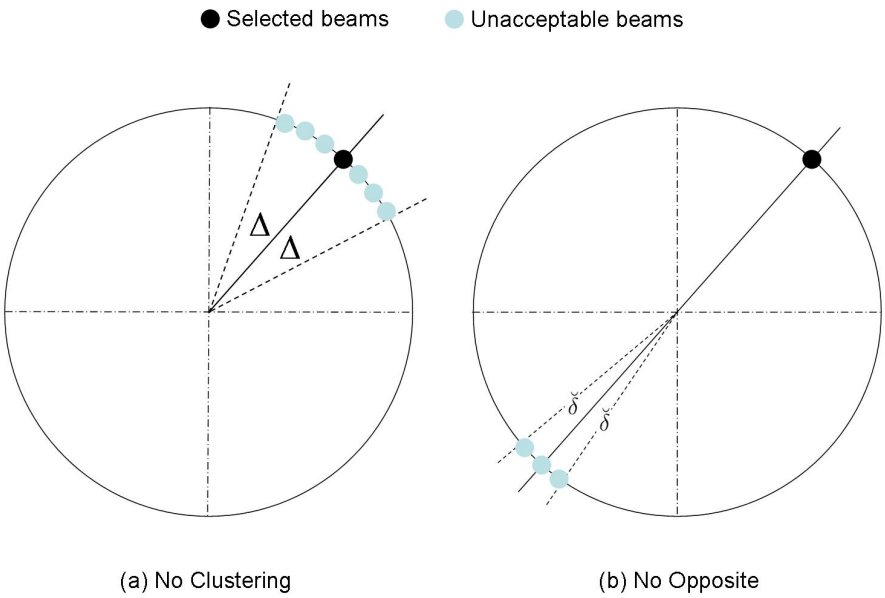


Figure 22: Constraints of beam-angle spacing: a. variables fixed to 0 by the minimum spacing constraints and b. variables fixed by the opposite spacing constraints.

$$w_{ij} \leq M_i x_i \tag{51}$$

where M_i is a positive constant that can be chosen as the largest possible intensity emitted from beam i.

Based on clinical experiments, some researchers provided beam-angle spacing constraints to prevent undesirable plans by considering geometric spacing among selected beams. For example, D'souza et. al's [35] has shown that directly opposed beams yield unexpected hot spots in the Corvus planning system. In our analysis, two types of beam-angle spacing constraints are used. One is for specifying the minimum spacing between geometrically adjacent beams. (See Figure 22. a) Suppose that Δ is the minimum angle of geometric spacing allowed between adjacent beams. The minimum spacing constraints can be modelled as follow:

$$\sum_{l=i}^{m+i} x_{l_{(\theta)}} \leq 1, \forall\ i \in B \tag{52}$$

where m is the largest number satisfying $m\delta \leq \Delta$ and $l_{(\theta)}$ is defined as: if $l\theta > 360^o$ then $l_{(\theta)} = l - \lfloor 360/\theta \rfloor$, otherwise $l_{(\theta)} = l$. Since the minimum spacing constraints do not allow the selection of adjacent beams within Δ^o, the resulting treatment plan is free from clustering of beams.

The other is to exclude directly (or nearly) opposed beams. As shown in Figure 22. b, if the difference of spacing angle between two beams is almost 180^o, both beams cannot be included in a treatment plan. Suppose that the opposite spacing angle is $180 \pm \breve{\delta}$. Then the constraints can be modelled as

$$x_i + x_{(i\breve{+}k)_{(\theta)}} \leq 1, \quad k = -\breve{m}, -\breve{m} - 1, ..., 0, ..., \breve{m}, \forall\ i \in \breve{B} \tag{53}$$

where $\breve{m}$ is the largest number satisfying $\breve{m}\theta \leq \breve{\delta}$, and $\breve{B}$ is the subset of B by taking all the indices in the first half space (0^o - 180^o). And $\breve{i}_{(\theta)}$ is defined to finding the index of beam angle directly opposed to beam i. That is,

$$\breve{i}_{(\theta)} := (i + \lfloor 180/\theta \rfloor)_{(\theta)}.$$

3.3 Solution Strategy: Branch-and-Price Method

The entire system for IMRT optimization problem $\mathcal{P}$ is:

$$\text{Min} \quad \gamma y^T Q y + cy$$

$$\begin{aligned}
\text{S.T.} \quad Dw - y &= b && (54)\\
Gy - Hz &= h && (55)\\
Ew - Fx &\leq 0 && (56)\\
\sum_{i \in B} x_i &\leq B_{max} && (57)\\
\sum_{l=i}^{m+i} x_{l_{(\theta)}} &\leq 1, \forall\ i \in B && (58)\\
x_i + x_{(i \breve{+} k)_{(\theta)}} &\leq 1, \quad k = -\breve{m}, -\breve{m}-1, ..., 0, ..., \breve{m}, \forall\ i \in \breve{B} && (59)\\
0 &\leq w\\
0 &\leq y\\
0 &\leq z\\
x &\in \{0, 1\}
\end{aligned}$$

where constraints (55) represent the constraints related to the dose-volume constraints, coverage constraints, and upper and lower bound constraints of PTV's and the critical structures. And constraint (56) represents equation (51). The above system has proved to be computational intractable, we will describe a branch-and-price method to solve these instances.

3.3.1 Branch-and-Price Method

Let Ξ denote the set of beams satisfying the beam selection constraint (50), the beam-angle spacing constraints (52) and (53). Note that Ξ is bounded. Let Ξ^* be the set

defined as $\Xi^* = \{x \in \Xi : x \ \ binary\}$. Then Ξ^* is made up of the finite number of points. Consequently, the binarity of x variables makes Ξ^* coincides with the extreme points of its convex hull, $conv(\Xi^*)$. That is, given $\Xi^* = \{\chi_1, ..., \chi_p\}$, any points $\chi \in \Xi^*$ can be represented by

$$\chi = \sum_{1 \le k \le p} \chi_k \lambda_k,$$

combined with following convexity constraint:

$$\sum_{1 \le k \le p} \lambda_k = 1, \text{ where } \lambda_k \in \{0, 1\}, \ k = 1, ..., p.$$

As representing a bounded polyhedron by the convex combination of its extreme points, Dantzig-Wolfe decomposition [26] is applied to the column generation approach to integer programming [7, 104]. Finally, the original problem $\mathcal{P}$ can be rewritten into its column generation form, called the master problem ($\mathcal{MP}$):

$$\text{Min} \qquad \gamma y^T Q y + cy$$

$$\begin{aligned}
\text{S.T.} \quad Dw - y &= b \\
Gy - Hz &= h \\
Ew - \sum_{1 \le k \le p} F\chi_k \lambda_k &\le 0 \\
\sum_{1 \le k \le p} \lambda_k &= 1 \\
\lambda_k &\in \{0, 1\}, \ k = 1, ..., p \\
0 &\le y \\
0 &\le w \\
0 &\le z.
\end{aligned} \tag{60}$$

Since there may be a huge number of generated columns in the column generation form, it is impossible to solve the instance of $\mathcal{MP}$ containing all these columns. Thus, a restricted instance, called a restricted master problem ($\mathcal{RMP}$), consisting of only a subset (i.e., $\Xi^*_{\mathcal{RMP}} = \{\chi_{\tilde{1}}, ..., \chi_{\tilde{p}}\}$) of its entire columns is considered in our column generation method. The general way to solve the mixed integer $\mathcal{MP}$ to optimality performs in a branch-and-bound framework. This is known as branch-and-price [7] or IP column generation [103]. In the structure of the branch-and-bound tree, a column generation method is used to solve the relaxation problem of $\mathcal{MP}$ at each node. If the solution to the relaxed $\mathcal{MP}$ is feasible to original problem, $\mathcal{P}$, or the lower bound of the relaxed $\mathcal{MP}$ is greater than the objective function value of the incumbent integral solution, the current node is pruned. Otherwise, branching will occur.

In the case when the relaxation problem of an $\mathcal{RMP}$ is LP, the reduced cost, $\bar{c}$, is calculated from the optimal solution to the relaxed problem and new column can be generated by solving the pricing problem:

$$\text{Min } \{\bar{c}\chi : \chi \in \Xi^*(\ or\ conv(\Xi^*))\}.$$

3.3.2 Column Generation

Since our relaxation problem is QP, we need how to determine the objective function of a pricing problem. The relaxation problem of the $\mathcal{RMP}$ is

$$\text{Min} \qquad \gamma y^T Q y + cy$$

$$\begin{aligned}
\text{S.T.} \quad Dw - y &= b && (61)\\
Gy - Hz &= h && (62)\\
Ew - \sum_{\tilde{1} \le k \le \tilde{p}} F\chi_k \lambda_k &\le 0 && (63)\\
\sum_{\tilde{1} \le k \le \tilde{p}} \lambda_k &= 1 && (64)\\
0 &\le \lambda\\
0 &\le y\\
0 &\le w\\
0 &\le z.
\end{aligned}$$

Let π_b, π_h, π_0, and π_1 denote the dual variables to the constraints, (61), (62), (63), and (64), respectively. Then the dual problem of the relaxed $\mathcal{RMP}$ is given as

$$\text{Max} \quad {\pi_b}^T b + {\pi_h}^T h + \pi_1 - \gamma \rho^T Q \rho$$

$$\begin{aligned}
\text{S.T.} \quad -{\pi_b}^T + {\pi_h}^T G - 2\gamma \rho^T Q &\le c && (65)\\
{\pi_b}^T D - {\pi_0}^T E &\le 0 && (66)\\
-{\pi_h}^T H &\le 0 && (67)\\
{\pi_0}^T F\chi_k + \pi_1 &\le 0, \quad \tilde{1} \le k \le \tilde{p} && (68)\\
\pi_0 &\ge 0\\
\pi_b,\ \pi_h,\ \pi_1,\ \text{and } \rho \quad &: \quad \text{unrestricted.}
\end{aligned}$$

Suppose that we have an optimal solution to the relaxed $\mathcal{RMP}$. That is, the solution satisfies all the constraints of its primal and dual problem. Clearly, it guarantees

the satisfaction of the constraints (68) for the subset, $\Xi^*_{\mathcal{RMP}}$ only. Thus we need to check if the solution is feasible to the entire set, Ξ^*. Similar to Dantzig-Wolfe decomposition methods applied to the linear programming, a pricing problem is introduced to find new entering columns with respect to the constraints (68):

$$\begin{aligned} \text{Min} \quad & -{\pi_0}^T F\chi \\ \text{S.T.} \quad & \chi \in \Xi^* \ (or, \ conv(\Xi^*)). \end{aligned} \tag{69}$$

Let $\tilde{\chi}$ be the optimal solution to the above sub-problem (69). If $\pi_1 > -{\pi_0}^T F\tilde{\chi}$, then $\tilde{\chi}$ is added into $\mathcal{RMP}$ and we proceed to solve the updated $\mathcal{RMP}$. In typical column generation, this iterative process is performed until $\pi_1 \leq -{\pi_0}^T F\tilde{\chi}$. From equation (51), the above subproblem (69) is equivalent to

$$\begin{aligned} \text{Min} \quad & \sum_i -\{M_i \sum_j {\pi_{0_i}}^j\}\chi^i \\ \text{S.T.} \quad & \chi \in \Xi^* \ (or, \ conv(\Xi^*)) \end{aligned}$$

where ${\pi_{0_i}}^j$ and χ^i denote the element of vectors, π_{0_i} and χ, respectively.

3.4 Computational Issues: Implementation

Since a dose coefficient matrix, D, is dense and large scale, it is proven to be difficult to solve the relaxation problem of the $\mathcal{RMP}$ when the columns generated at a node cover many beam variables, x_i. Table 10 shows the solution times of the pure QP instances of the relaxation problem obtained by the commercial Cplex® barrier optimizer. The table clearly illustrates the difficulty of solving the relaxation problem itself as well as the dramatic increase in solution times with addition of only a few beam variables to $\mathcal{RMP}$. Thus solving the IMRT instances are proven to be very difficult when there are a huge set of beam variables, since a typical branch-and-price approach requires the solutions to a large number of $\mathcal{RMP}$ relaxations.

Table 10: Solution times of the QP relaxation of $\mathcal{RMP}$ by the barrier optimizer on Cplex®

# beams	Instance size			Cplex®
	# rows	# columns	# nonzeros	CPU Times (Sec)
8	26538	11286	792399	9873
10	26845	11595	989907	12182
12	27219	11971	1198756	18933

3.4.1 Restricted Branch-and-Price Method

We propose and investigate a column generation scheme with limited number of original binary variables for the beam angles. *We call this scheme the restricted branch-and-price method.* This approach begins with partitioning the binary variables, x, into two subsets: the *Active Set*, AS, and the *Inactive Set*, IS. While all the variables in IS are fixed to 0, only variables in AS are allowed to set be set to 0 or 1. And when a variable in AS is fixed to 0 by branching (*or cutting plane*), a variable in IS can be moved to AS set. Thus we maintain the size of *Active Set* to a certain threshold, κ: $|AS| \leq \kappa$.

Thus, since $\mathcal{RMP}$ consists of the columns generated from at most κ of the original binary variables, the computational burden in solving the relaxed $\mathcal{RMP}$ can be monitored to make the solution time practical. Moreover, at each node of the branch-and-bound tree, column generation can be terminated earlier by stopping when no column combined with the variables in $\{AS \cup IS_\kappa\}$ is generated from the pricing problem. Note that IS_κ is determined at the column generation as:

$$IS_\kappa \subseteq IS \text{ and } |IS_\kappa| \leq \kappa - |AS|.$$

This can also prevent the *tailing-off effect* which makes column generation algorithms difficult to proceed, as it drags the process through a large number of iterations before proven optimality is achieved in the relaxation problem.

In terms of the practical solution time for IMRT treatment planning, an additional

advantage results from every node consisting of a relatively small number of beam variables in the branch-and-bound framework: in the mid of the solution process, we can get a feasible solution easily (for actual clinical usage) by a few branching steps of beam variables that are in the active set. Thus this active set approach can also be viewed as a "clinically adaptive" approach.

► Algorithm: Restricted Branch-and-Price for IMRT

Step 1. Initialize the best IP feasible solution, $z^\star$.

Step 2. Determine initial active set AS and inactive set IS and generate initial $\mathcal{RMP}$ with AS.

Step 3. Solve the relaxation problem of current $\mathcal{RMP}$.

Step 4. Update the best IP solution, $z^\star$, by this current solution if solution is integral feasible, and whose optimal value is better than the current best IP solution.

Step 5. Compute the lower bound, $\underline{z}$. If $\underline{z} > z^\star$, then go to Step 8.

Step 6. Run the heuristic column generation algorithm with the *reduced costs*, $-M_i \sum_j \pi_{0_i}^j$, computed from the dual solution to the relaxation of $\mathcal{RMP}$. If columns are generated, add them to $\mathcal{RMP}$ and go to Step 3.

Step 7. If $|AS| > B_{max}$, then do branching (on the active set), update the node list, and (*update* $\mathcal{RMP}$) and go to Step 3.

Step 8. Prune the node. If there exist a node unconquered at the node list, then (*update* $\mathcal{RMP}$) and go to Step 3.

Step 9. Stop.

◄

3.4.2 Initialization of AS and IS

Since the objective function of the pricing problem composes of the dual solution to the relaxation problem $\mathcal{RMP}$, we need to determine an initial $\mathcal{RMP}$ at the beginning of the branch-and-price scheme. Obviously, the initial $\mathcal{RMP}$ must be feasible to its relaxation problem so as to allow one to formulate the proper pricing problem. And the "better" the initial $\mathcal{RMP}$ is chosen, the better the performance of the overall process will be.

In the applications to IMRT, our procedure starts with determining an initial Active Set, AS, of beam orientations. Typically, it may be selected as a set of beams with equi-spaced angles (i.e., 0, 40, 80, 120, 160, 200, 240, 280, and 320°) within the predetermined κ. A better collection of beams can be obtained by considering the geometric information of PTV's and critical structures based on clinical experience. For example in the conventional conformal radiotherapy, the beam orientations are selected according to the beam's-eye view of PTV, which offers a good selection of beams targeting the PTV while avoiding direct path to some critical structures.

3.4.3 Heuristic Algorithm for Column Generation

As mentioned in the previous section, we need to generate new columns with limited number of original binary variables partly because of the difficulty of $\mathcal{RMP}$, and partly to maintain feasible and clinically acceptable plans during the solution process. Therefore, we include the following constraint in the pricing problem:

$$\sum_{i \in IS} x_i \leq \kappa - |AS|. \tag{70}$$

Note that since our $\mathcal{RMP}$ covers only variables in AS, we need to estimate the dual variables, π_0, corresponding to the constraints related to IS. From equations (51) and (66), we have

$$\pi^j_{0_i} \geq D^T_j \pi_b.$$

where D_j is the j^{th} column vector of dose matrix, D. Thus, at i^j constraint related to IS, the value of a dual variable, $\pi^j_{0_i}$, can be set to $D^T_j \pi_b$.

Clearly, it is not easy to solve the pricing problem with binary variables. To generate new column with the limited number of original binary variables, we derived a heuristic method to find promising columns at a node based on the geometric interpretation of spacing constraints.

Suppose that AS^r and IS^r be the active set and inactive set at current node r, respectively. Let $AS(1)^r$ be the subsets of AS^r fixed to '1' by branching decisions. Let d_i denote $-\{M_i \sum_j \pi^j_{0_i}\}$. As soon as solving $\mathcal{RMP}$ at the node r, we can calculate d_i for each i. Given d_i, for $i \in \{AS^r \setminus AS^r(1) \cup IS^r\}$, the heuristic algorithm for column generation is following:

- Algorithm: Heuristic column generation

Step 1. Initialize new column index set, $I^r \leftarrow AS^r(1)$.

Step 2. Make the index set, I, whose elements are listed in the smallest order of d_i

Step 3. If $|I^r \cap IS^r| = \kappa - |AS^r|$, update $I \leftarrow I \backslash IS^r$.

Step 4. If $I = 0$, then return '*no column found*' message. Otherwise, choose the first index, $I_{(1)}$, in I. And update $I^r \leftarrow I^r \cup \{I_{(1)}\}$ and $I \leftarrow I \setminus \{I_{(1)}\}$

Step 5. Delete unacceptable indices in I based on the constraints of geometric spacing, (52) and (53). That is, as selecting a variable, $x_{I_{(1)}}$, the corresponding variables to adjacent beams within $\pm\Delta^o$ are fixed to 0 by the minimum spacing constraints, (52). (See Figure 22. a) And the corresponding variables to opposed beams $(180^o \pm \breve{\delta})$ are also fixed to 0 by the opposite spacing constraints, (53). (See Figure 22. b)

Step 6. If $|I^r| < B_{max}$, go to step 3.

Step 7. If $\sum_{I^r} d_i + \pi_1 > 0$, then update $AS^r \leftarrow AS^r \cup I^r$ and $IS^r \leftarrow IS^r \backslash I^r$, and return I^r as the index set of new column. Otherwise, return '*no column found*' message.

Step 8. Stop.

◀

3.4.4 Branching

For general mixed integer $\mathcal{MP}$s, one of the typical solution strategies is to branch on the original integer variables so to be feasible to their integrality [7, 44]. Suppose the solution to the relaxation of current $\mathcal{RMP}$ $\bar{\chi}$ has a fractional component, i, clearly, $\bar{\chi}$ is infeasible to the original problem $\mathcal{P}$. Thus, we have to do branching with the following rule: one branch is to

$$\sum_{1 \le k \le p^i} \chi_k^i \lambda_k^i \le \lfloor \bar{\chi}^i \rfloor$$

and the other branch is to

$$\sum_{1 \le k \le p^i} \chi_k^i \lambda_k^i \ge \lceil \bar{\chi}^i \rceil.$$

Note that the component i is the element of the active set. In our experiments on IMRT, when there exist several fractional components on $\bar{\chi}$ we choose the closest component to 0. As the component i is fixed after branching, we need to update AS^{r+1} and IS^{r+1} of its children. Clearly, the active set may not change at the child node derived from the latter branch, $\lceil \bar{\chi}^i \rceil$. Since this makes the redundant node to its parents node, we will do branch recursively in this case. Suppose that $AS^r \setminus AS^r(1)$ is not empty. Given $\bar{\chi}^i, i \in AS^r \setminus AS^r(1)$, the child nodes are generated as following:

▶ Algorithm: branching

Step 1. Select i by $argmin\{\bar{\chi}^i, i \in AS^r \setminus AS^r(1)\}$.

Step 2. For the former branch, $\lfloor \bar{\chi}^i \rfloor$, we update that $AS^{r+1} \leftarrow AS^r \setminus \{i\}$.

Step 3. For the latter, $\lceil \bar{\chi}^i \rceil$, we update the child as following:

- Check if beam i is conflicted with any beam in $AS^r(1)$, by similar way to Step 5 in Algorithm: Heuristic column generation. If the beam i is infeasible, this child node is pruned. Otherwise, update $AS^{r+1}(1) \leftarrow AS^r(1) \cup \{i\}$.
- If $AS^{r+1}(1) \geq B_{max}$, then set $AS^{r+1} \setminus AS^{r+1}(1) \leftarrow 0$ and $IS^{r+1} \leftarrow 0$ and goto Step 4.
- Update AS^{r+1} and IS^{r+1} by eliminating all beams in $\{AS^r \setminus AS^r(1) \cup IS^r\}$ which are conflicted with beam i. If AS^{r+1} is identical to AS^r, set $r \leftarrow r + 1$ and goto Step 1.

Step 4. Stop.

◀

3.4.5 Lower Bound and Early Termination

In a branch-and-bound tree in which branch-and-price algorithms perform, it is advantageous to find strong bounds at each node. The stronger the lower bound, LB, the earlier the termination of the tree process: if LB is greater than the incumbent integral solution value, then one can prune the current node, else if LB is greater than the upper bound of $\mathcal{MP}$, then we can stop column generation. Moreover, this is very important in terms of controlling the *tailing-off effect* [105]. However, there is a trade-off in terms of the computational efforts. In generally it takes much computational time to obtain a strong bound.

Proposition: Given a dual solution, π, to $\mathcal{RMP}$, the lower bound of the relaxation problem of $\mathcal{MP}$, z^r_{LP}, at node r is

$$z^r_{LP} \geq {\pi_b}^T b + {\pi_h}^T h + -\gamma \rho^T Q \rho + min_{x \in \Xi^*(r)} \sum_i -\{M_i \sum_j {\pi_0}_i{}^j\} x^i.$$

Proof: At node r, the lagrangian relaxation of $\mathcal{P}$ gives

$$\begin{aligned} z^r_{LP} &\geq min \left\{ \begin{array}{c} x \in \Xi^*(r) \\ w \geq 0 \\ y \geq 0 \\ z \geq 0 \end{array} \right\} \left\{ \begin{array}{l} \gamma y^T Q y + cy + \pi_b^T (b - Dw + y) \\ +\pi_h^T (h - Gy + Hz) + \pi_0^T (Ew - Fx) \end{array} \right\} \\ &= {\pi_b}^T b + {\pi_h}^T h + min \left\{ \begin{array}{c} x \in \Xi^*(r) \\ w \geq 0 \\ y \geq 0 \\ z \geq 0 \end{array} \right\} \left\{ \begin{array}{l} \gamma y^T Q y + (c + \pi_b^T - \pi_h^T G) y \\ -(\pi_b^T D - \pi_0^T E) w + \pi_h^T H z - \pi_0^T F x \end{array} \right\}. \end{aligned}$$

Since π's is an optimal solution to $\mathcal{RMP}$, from the dual problem of $\mathcal{RMP}$, we have that for $w \geq 0,\ y \geq 0,\ z \geq 0$,

$$-{\pi_b}^T y + {\pi_h}^T G y - 2\gamma \rho^T Q y \leq cy \tag{71}$$

$${\pi_b}^T D w - {\pi_0}^T E w \leq 0 \tag{72}$$

$$-{\pi_h}^T H z \leq 0. \tag{73}$$

Therefore, the lagrangian relaxation is bounded by

$$z_{LP}^r \geq {\pi_b}^T b + {\pi_h}^T h + min \left\{ \begin{array}{c} \\ x \in \Xi^*(r) \\ \\ w \geq 0 \\ \\ y \geq 0 \\ \\ z \geq 0 \\ \\ \end{array} \right\} \{\gamma y^T Q y - 2\gamma \rho^T Q y - \pi_0^T F x\}.$$

Since Q is positive definite and γ is positive, we also have that

$$\begin{aligned} 0 &\leq \gamma(y-\rho)^T Q(y-\rho) \\ &= \gamma y^T Q y - 2\gamma \rho^T Q y + \gamma \rho^T Q \rho \\ &\Updownarrow \\ -\gamma \rho^T Q \rho &\leq \gamma y^T Q y - 2\gamma \rho^T Q y. \end{aligned} \tag{74}$$

This means that

$$z_{LP}^r \geq {\pi_b}^T b + {\pi_h}^T h + -\gamma \rho^T Q \rho + min_{x \in \Xi^*(r)} \{-\pi_0^T F x\}$$

Finally, from the definition of coefficient matrix F, we can get the lower bound of the relaxation problem. □

Based on the above Proposition, we have derived the lower bound of the relaxed $\mathcal{MP}$ at each node from the solution to the $\mathcal{RMP}$ relaxation by column generation with the limited number of original binary variables. That is, if inactive set, IS^r, is empty at current node r, we can get exact solution of $min_{x \in \Xi^*(r)} \sum_i -\{M_i \sum_j {\pi_0}_i{}^j\}$ by enumerating the feasible set of beam angles in active set, AS^r; otherwise, we obtain the minimum value heuristically by taking B_{max} number of beam variables in smallest order of reduced cost, d_i.

3.4.6 Solving $\mathcal{RMP}$

The main purpose of solving the relaxation problem of $\mathcal{RMP}$ is to provide a *good* dual solution which composes the objective function of the pricing problem.

In the case when there exists a optimal face defined by alternative dual solutions, it is believed that the interior point methods return a better solution than the simplex algorithm. Because an interior point algorithm searches a point in the relative interior of the face, whereas the simplex algorithm enumerates and moves from one extreme point to another. Further, computational tests have proved that $\mathcal{RMP}$ is far more tractable using our in-house interior-point algorithm. Therefore, we applied our in-house interior point algorithm, developed based on the concept from the predictor-corrector algorithm [61, 62, 69, 82] to solve $\mathcal{RMP}$. Lustig et al's work is considered one of the most popular interior-point algorithm to-date to solve large-scale linear programming problems.

Despite the objective function of the relaxed $\mathcal{RMP}$ is quadratic, since Q is a diagonal matrix, the predictor-corrector algorithm can be readily applied to the relaxation problem. Our QP algorithm described in Chapter 2 showed its efficiency based on real clinical and patient cases. To improve the solution time within acceptable tolerance of the solution quality, we also incorporate the matrix reduction scheme that we have described in Chapter 2 in solving $\mathcal{RMP}$. In some senses, the devastation resulting from reduced coefficient may be negligible. Further, the plan quality can also be improved and achieved by the post-processing procedure described in Chapter 2.

3.5 Computational and Clinical Results

3.5.1 Patient Data and Instance Generation

We use the Computational Environment for Radiotherapy Research (CERR) software [31] to read in real patient images and generate the patient information and dose data for treatment planning. CERR®, implemented Deasy's group at Washington University in St. Louis, works on Matlab® environment and provides various functional components including the routines of importing and visualizing the AAPM/RTOG treatment planning format, the contouring tools to edit and/or create anatomical structures, the routines of generating the influence matrix with a beam setup, and the plan analysis tools to generate and visualize DVH graphes, iso-dose curves, and dose color-washes from the solution to treatment planning. In particular, the influence matrix of each patient was calculated with the Quadrant Infinite Beam (QIB) algorithm [46] which is a fast method of 3-D dose calculation based on Ahnesjos approximation of kernel dose deposition patterns. Note that the influence matrix consisting of primary dose as well as scatter and results in about 10% of density in the coefficient matrix.

In our computational experiments, five patient cases including three Head-and-Neck cases and two prostate cases were used. The beam orientations are discretized using 10° over the gantry angle space and these beams were considered as the set of potential candidate beams. The patient information and dose data involves are very large scale and cannot be handled manually or interactively (as it takes too much computational time). Moreover, we need additional information to determine the instance parameters for IMRT treatment planning problems. Thus, we developed several computer modules to support the modelling process, including

- Extracting the geometry information of beams and selecting a set of good candidate beams

- Generating the dose matrix of each beam and extracting the spatial and organ relational information
- Assembling all data and formulating the MIQP instances
- Solving the instances with various solution techniques
- Analyzing the quality of solution plan using various clinical metrics.

3.5.1.1 Prostate cancer

Prostate cancer is very common disease in older men in the United States, in particular, most prevalent in men over age 65. Recent statistics shows the incidence of diagnosed prostate cancer has increased dramatically among American men.

Table 11 describes two prostate cases. And Table 12 summarizes the constraint parameters for the beamlet intensity map optimization. As shown in both tables, the prostate cases contain primary and secondary PTVs whose prescription doses are specified at 75.6 Gy and 45 Gy, respectively. For their coverage constraints, at least 95 % of each PTV should receive its prescription dose. In addition to critical structures, a *5-15mm critical-normal-tissue-ring* whose voxels are selected from normal tissues will be considered as one of the critical structures. For the rectum and the bladder, the DV constraints specify that 70 % of the rectum and the bladder should receive less than 70 and 80 Gy, respectively.

Table 11: Volume of each structure for the Prostate cases

Structures	case1	case2
	volume (cm^3)	
PTV 1	118.54	102.18
PTV 2	182.35	38.90
Bladder	98.32	148.18
Rectum	67.01	104.47
Left Femur head	0	162.04
Right Femur head	0	170.01
Unspecified Tissue	8072.11	10597.36

Table 12: Constraint parameters for the beamlet intensity map at Prostate case

Structures	Prdose (Gy)	Low Bd (Gy)	Upper Bd (Gy)	DVH
PTV 1	75.6	70.46	87.05	95% $\geq$ 75.6 Gy
PTV 2	45	44.25	87.05	95% $\geq$ 45 Gy
Bladder	0	0	87.05	70 % $\leq$ 80 Gy
Rectum	0	0	87.05	70 % $\leq$ 70 Gy
Left Femur head	0	0	87.05	
Right Femur head	0	0	87.05	
Unspecified Tissue	0	0	87.05	

3.5.1.2 Head-and-Neck cancer

Head-and-neck cancer refers to a range of malignant tumors located in or around the throat, larynx, nose, sinuses, and mouth. Hence, it affects seriously the vital functions such as eating and breathing. According to 2007 report from American Cancer Society, head-and-neck cancer accounts for about 3% to 5% of all cancers in the United States. And it is estimated that about 45,660 people will develop head-and-neck cancer, and about 11,210 will die from it. Table 13 describes three head-and-neck cases, and Table 14 shows the clinical dose constraint parameters.

Table 13: Volume of each structure of the Head-and-Neck cases

Structures	case1	case2	case3
	volume (cm^3)		
PTV 1	266.39	253.34	105.3107
PTV 2	0	116.62	38.7295
PTV 3	430.64	285.47	522.3822
Spinal cord	20.66	24.35	9.9201
Brain stem	29.33	13.98	21.3629
Right Parotid	36.18	40.39	56.5862
Left Parotid	24.08	0	61.8214
Right Submandibular	20.59	0	0
Left Submandibular	0	0	9.4446
Unspecified Tissue	2666.93	3543.48	10954.21

Table 14: Constraint parameters for the beamlet intensity map at Head-and-Neck case

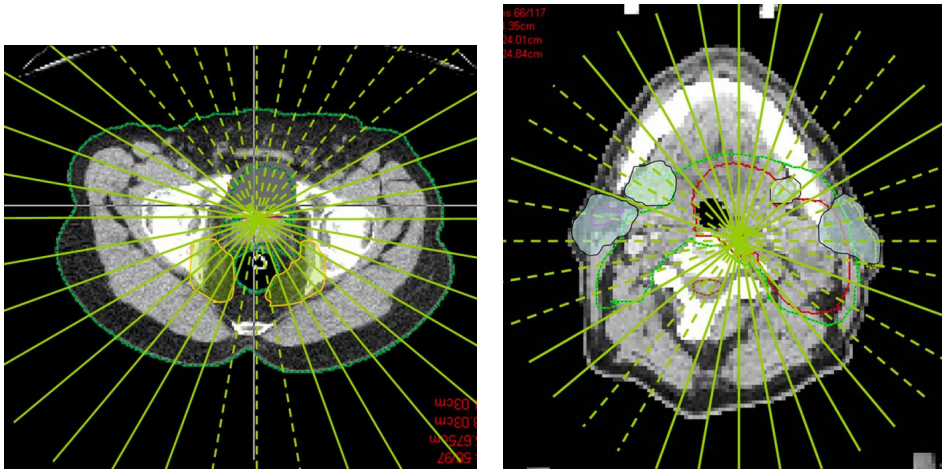
Structures	Prdose (Gy)	Low Bd (Gy)	Upper Bd (Gy)	DVH
PTV 1	72	68.75	83.8	95% $\geq$ 72 Gy
PTV 2	54	52.65	83.8	95% $\geq$ 54 Gy
PTV 3	49.5	48.35	83.8	95% $\geq$ 49.5 Gy
Spinal cord	0	0	45	
Brain stem	0	0	54	
Right Parotid	0	0	76.32	50% $\leq$ 30
Left Parotid	0	0	76.32	50% $\leq$ 30
Right Submandibular	0	0	76.32	50% $\leq$ 30
Left Submandibular	0	0	76.32	50% $\leq$ 30
Unspecified Tissue	0	0	83.8	

3.5.1.3 Determination of Beam Selection Parameters

In additions to dose/dose volume constraints, we will determine three types of parameters for the constraints associated with the beam angle selection as described at section 3.2.4. Table 15 shows the parameter values of the maximum number of beam angles allowed in any feasible plans and the beam-angle spacing constraints. Note that '$\Delta = 30°$' means that the minimum angle between two adjacent beams is 30° and '$\breve{\delta} = 0$' implies that two beams opposed directly cannot be selected at the same time. As a result, we selected two sets of candidate beams, whose sizes are 15 and 20, based on the relationship of geometry between each beam and the anatomical structures. Figure 23 illustrates the selected beams whose central rays escaping the critical structures.

Table 15: Predetermined model parameters related to the selection of beam orientations and beam setups

	values
B_{max}	7
Δ	30°
$\breve{\delta}$	0°
κ	8, 9, 10
# candidate beams	15, 20



(a) Prostate case (b) Head-and-Neck case

Figure 23: Selection of candidate beams. Each indicates the central ray of beam. Only solid line escaping the critical can be selected as a candidate beam.

Moreover, it is very important to establish the proper value of κ, the maximum size of a active set, in restricting branch-and-price process and in maintaining each feasible solution to be clinically acceptable. Further, this is critical also as the solution time of the relaxation problem of $\mathcal{RMP}$ depends on the size of the active set. For computational comparison, we determined three different values of κ and analyzed how the size of κ affects the overall solution times.

3.5.2 Computational Results

The computational experiments have been carried out with the our "in-house" optimization code based on the interior point algorithm and the branch-and-price algorithm prescribed at the previous section. The following numerical results were obtained when running our solvers on a 2.33 GHz Intel Xeon machine with 12 GB of RAM.

As mentioned in previous sections, our dose matrix is large scale and relatively dense. The size and density depend on which structures are included into the model, and how the dose distribution is computed from the beamlets to voxels. In this

research, we first observed that critical ring structures and scatter doses can be negligible in the selection of beam orientations. Based on these observations, the dose matrices for the optimization problem of the beam angle selections can be reduced into only primary doses computed from PTVs and the critical structures. This reduction improves drastically the computational time in solving relaxation of RMP via the column generation method.

3.5.2.1 Optimization of Beam Orientations

Starting with the initial setups, we created MIQP instances and solved them by restricted branch-and-price algorithm keeping the limited number of beams in the active set. Consequently, we have six test-runs associated with the size of the active set and the number of candidate beams at each patient case. Table 16 and 17 summarize the computational statistics for each patient case. The third and fifth columns show the total # QP and total # iterations used to solve the relaxation of QP in the entire branch-and-price process, respectively. The last two columns show the CPU time elapsed and the objective value of the optimal IP solution, respectively.

The computational results of both patient cases show similar trends to each other. The solution time of active set = 8 is much faster than the other two cases, 9 and 10. Since increasing a few candidate beams may provide a large numbers of feasible beam combinations, the solution times from 20 candidate beams are far much longer than the ones from 15 beams.

We remark that the head-and-neck cases have much larger PTVs than the prostate cases, thus result in much larger instances of $\mathcal{RMP}$. As a result, every solution time is longer than those from the prostate cases.

We have also investigated some scheme of "warmstart" which may provide a good initial solution for the child node within a branch-and-bound framework. From the computational experiments of Lee and Mitchell [58], their "advanced warmstart"

showed the efficient performance on solving the instances of MIPLIB and the portfolio problem. Unlike typical branch-and-bound algorithms, our branch-and-price algorithm adds (or delete) hundreds of dense columns corresponding to each entering (or eliminating) beam at the child RMP. This causes a dramatic change of the child node from its parent node. Consequently, the warm-start scheme applied to our IMRT instances did not provide much solution time improvement at the series of interior point iterations.

Table 16: Computational statistics at Prostate cases

Size of	#candidate	Relaxation of QP		IP Solution	
Active Set	beams	#QP	#iterations	CPU sec.	Obj. Val.
case1					
8	15	45	984	6911	1147.3
	20	232	5059	34378	1144
9	15	197	4301	37793	1147.3
	20	1229	26979	219278	1144
10	15	187	4079	44928	1147.3
	20	1578	34629	278055	1144
case2					
8	15	55	1545	967	1288.5
	20	248	6812	4025	1263.3
9	15	174	4689	3438	1288.5
	20	1311	34410	24379	1263.3
10	15	255	6935	5623	1288.5
	20	1580	40944	29972	1263.3

Table 17: Computational statistics at Head-and-Neck cases

Size of Active Set	#candidate beams	Relaxation of QP #QP	#iterations	IP Solution CPU sec.	Obj. Val.
case1					
8	15	57	1511	10145	901.75
	20	372	10263	60874	901.75
9	15	159	4211	38661	901.75
	20	1499	40094	341585	901.75
10	15	185	4846	55963	901.75
	20	1876	49882	460161	901.75
case2					
8	15	64	1548	8651	166.12
	20	393	9785	50254	161.11
9	15	159	3804	20629	166.12
	20	1659	40837	219041	161.11
10	15	232	5505	41248	166.12
	20	1798	43767	256555	161.11
case3					
8	15	137	7001	15884	498.7
	20	866	40128	91001	492.76
9	15	218	8699	26445	498.7
	20	1965	67482	196461	492.76
10	15	276	10436	29282	498.7
	20	2247	74121	221983	492.76

3.5.2.2 Optimization of Beam Intensity Map

In second phase, we optimized the beamlet intensity map with the full dose matrix associated with the set of beam orientations obtained from the first phase. Table 18 shows the size of the instances generated from the optimal set of beams for each run and the CPU time between our QP solver and Cplex® barrier algorithms.

3.5.3 Clinical Results

We presented here the treatment plans for one prostate case and one head-and-neck case. Both have been processed using the optimal solutions obtained with 15 candidate beams. The solution plans are called the *Column Generation Plan*s (CG-P).

Table 18: Computational statistics for beamlet intensity map. Here, Error 1001 means "CPLEX Error 1001: Out of memory".

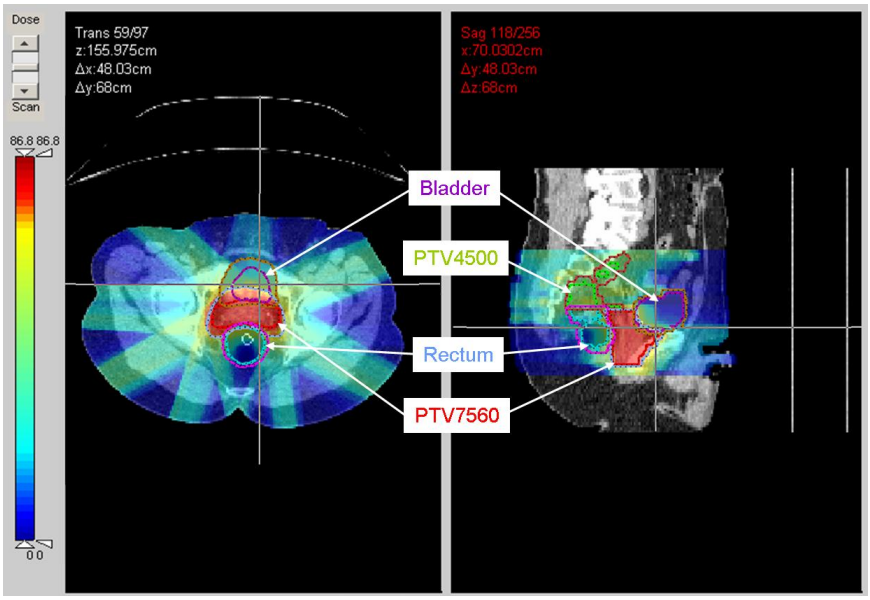
Patient cases	#cand. beams	Size of instances (preprocessed) #nonzeros	#rows	#cols (#dense)	CPU Sec QP SOL	Cplex®
Prostate						
case1	15	2525130	38640	11195 (1167)	352	85396
	20	2540397	38640	11197 (1169)	379	96130
case2	15	1880233	37389	17499 (589)	189	41673
	20	1873525	37445	17496 (587)	192	41733
Head-and-Neck						
case1	15	3805826	65770	22632 (1649)	783	Error 1001
	20	3805826	65770	22632 (1649)	783	Error 1001
case2	15	4004531	69820	16979 (1330)	975	Error 1001
	20	4029571	69819	16986 (1337)	972	Error 1001
case3	15	3924922	83157	17952 (1227)	688	Error 1001
	20	3940218	83667	18092 (1253)	576	Error 1001

Each plan is evaluated by graphical and statistical tools such as dose-washes, dose-volume histogram, and statistical indices.

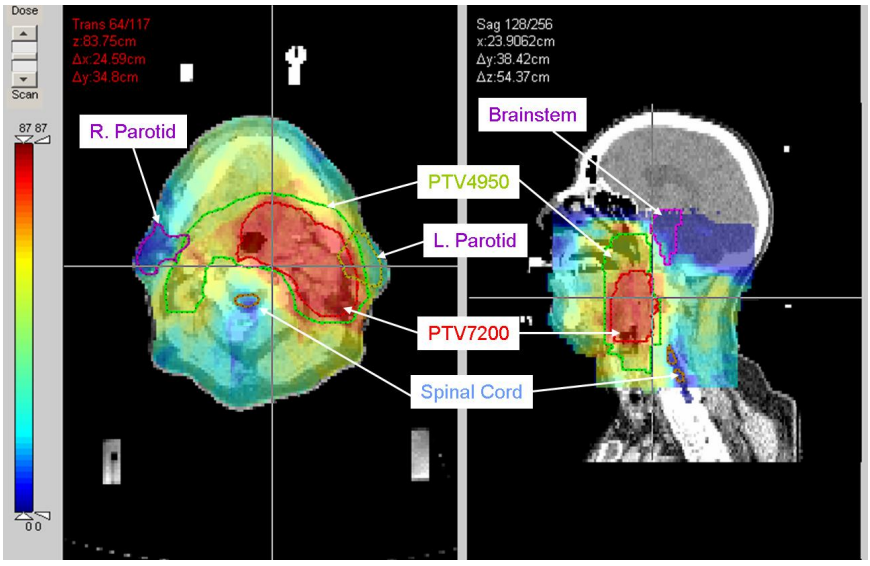
Figure 24 shows the dose-washes on the transverse view and sagittal view for our column generation plan. A dose-wash picture describes the overall dose distribution of the plan by scaled colors. Note that the color ranges from dark blue to dark red indicate the dose values from 0 to the maximum dose. The picture of prostate case (Figure 24 (a)) illustrates that most of rectum and bladder receive very low doses and the higher doses of 75.6 Gy or greater cluster around the PTV7560 region, between on bladder and rectum, and the PTV4500 placed above the rectum. Further, PTV7560 is covered by its prescription dose. In the head-and-neck case (Figure 24 (b)), the treatment plan delivers relatively high doses to some normal tissues because of the large PTVs that overlap with some of the critical structures. However, the plan spares all the critical structures and achieves full coverage of the prescription dose to PTVs.

As described at section 3.2.3, a Dose-Volume Histogram depicts the relationship between a dose level and the fractional volume of structure exposed greater than the particular dose: for a series of dose levels represented on X-axis, the corresponding

fractional values represented on Y-axis take the shape of curve. Note that a better treatment plan makes vertical curves over the prescription doses for PTVs as well as



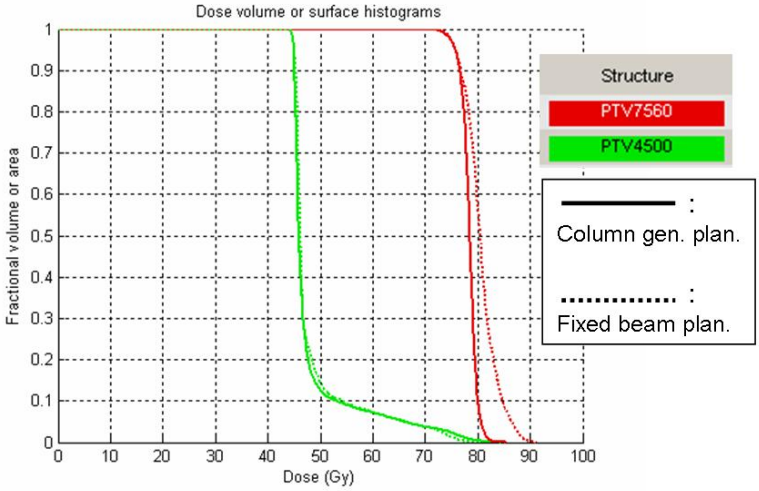
(a) Prostate case



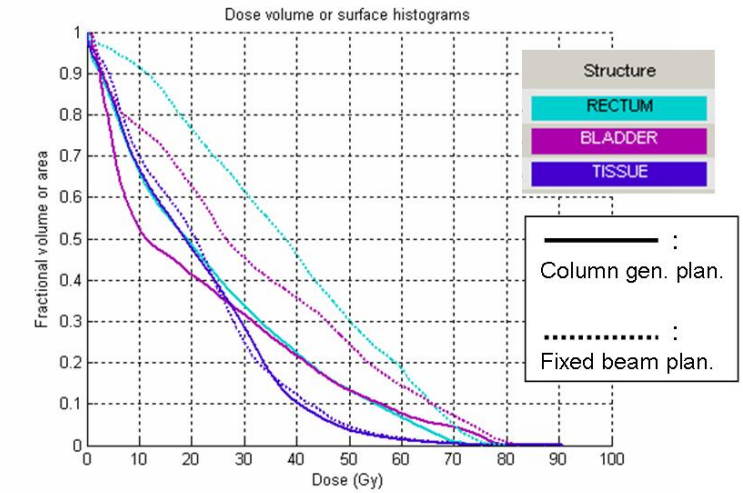
(b) Head-and-Neck case

Figure 24: Dose-washes on a transverse and a sagittal view generated from the optimal column generation plans. Scaled color bar indicates the various levels of dose value.

curves cluster low toward the Y-axis and X-axis for critical structures and normal tissues. Moreover, by plotting several treatment plans at one graph, one can compare the quality of the resulting plans. Figure 25 and 26 show the DVH curves comparing the optimal column generation plan with the fixed beam plan (from clinic) for prostate



(a) PTV7560 and PTV4500

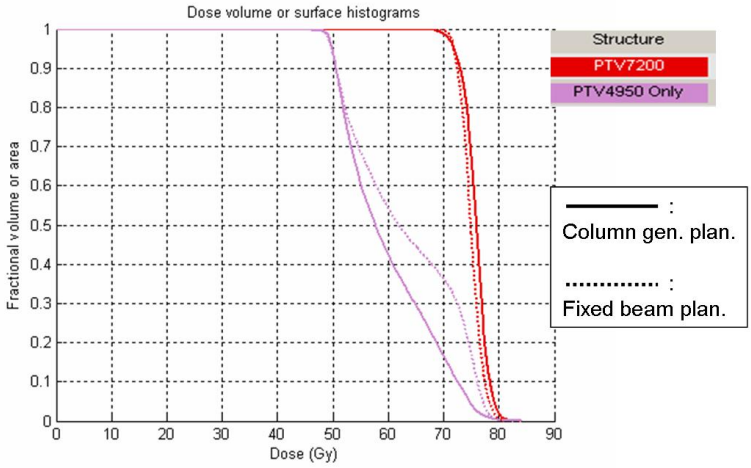


(b) critical structures and normal tissue

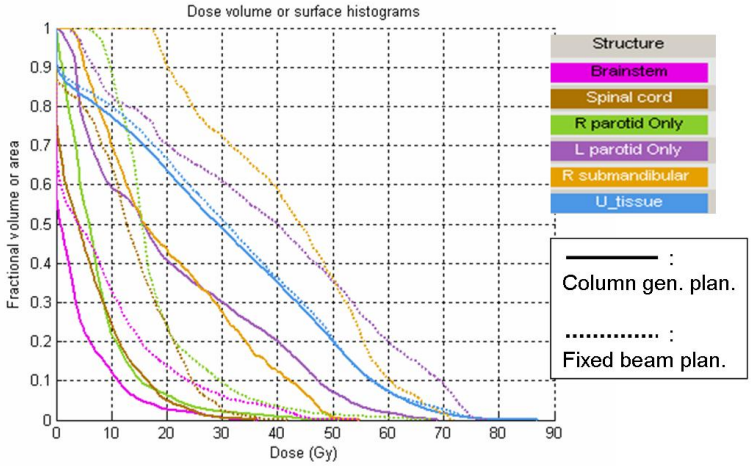
Figure 25: Dose-volume histograms at Prostate case. Note that the fixed beam plan (clinical plan) was obtained with 7 equi-spaced beams.

case and head-and-neck case, respectively. Here the *Fixed Beam Plan*, denoted as FB-P, was obtained by solving the quadratic programming problems based on the dose-based objective function with the 7 equi-spaced angles provided by the clinicians.

In addition to maximum, mean and minimum doses, several clinical metrics such as coverage index, homogeneity index, conformity index, and V_α values are computed to evaluate the treatment plans. The former three indices are used to the evaluation



(a) PTV7200 and PTV4950



(b) critical structures and normal tissue

Figure 26: Dose-volume histograms at Head-and-Neck case. Note that fixed beam plan (clinical plan) was obtained with 7 equi-spaced beams.

of PTV according to the Radiation Therapy Oncology Group (RTOG) guidelines. Recall that homogeneity index was described in section 3.2.2. Coverage is computed as the ratio of the target volume enclosed by the prescription isodose surface to the total target volume. Thus the coverage of a plan is always less than or equal to 1. Conformity is a measure of how well the prescription isodose surface conforms to the target volume. That is, it is computed as the ratio of the total volume enclosed by the prescription isodose surface to the target volume enclosed by this same surface. Conformity is always greater than or equal to 1. And as described at section 3.2.3, V_α is the fractional volume of a structure receiving greater than or equal to α dose. For critical structures, the lower the V_α, the better the treatment plan. Table 19 - 22 contrasted the clinical metrics of the resulting plans.

Clearly, all metrics show that our column generation plans are much superior than the fixed beam plans: column generation plan achieved more conformal plan to PTV with less exposure to the critical structures. Specifically, 70% of rectum and bladder received less than 40 Gy. And for the dose-volume constraints for the left parotid and the right submandibular, each prescribed to receive less than 30 Gy, the column generation plan spares more than 46% of left parotid and 31% of right submandibular than the fixed beam plan.

Table 19: DVH statistics of PTVs at Prostate case

	PTV7560	PTV4500
	CG-P/FB-P	CG-P/FB-P
Max (Gy)	85.45/91.85	84.35/82.85
Mean (Gy)	78.45/81.87	48.35/48.90
Min (Gy)	69.45/70.05	44.05/44.10
Coverage	**0.957/0.959**	**0.962/0.974**
Homogeneity	**1.23/1.31**	
Conformity	**1.25/1.48**	

Table 20: DVH statistics of critical structures and normal tissues at Prostate case

	Rectum	Bladder	Tissue
	CG-P/FB-P	CG-P/FB-P	CG-P/FB-P
Max (Gy)	75.45/81.95	82.45/84.55	90.75/89.05
Mean (Gy)	**23.90/42.72**	**22.05/34.55**	**20.62/21.29**
Min (Gy)	0.15/0.55	0.15/0.45	0.05/0.05
V_{20}	0.487/0.766	0.414/0.624	0.475/0.523
V_{40}	0.225/0.460	0.218/0.358	0.104/0.121
V_{60}	0.068/0.189	0.078/0.143	0.013/0.014

Table 21: DVH statistics of PTVs at Head-and-Neck case

	PTV 7200 CG-P/FB-P	PTV 4950 CG-P/FB-P
Max (Gy)	83.95/81.05	84.05/82.65
Mean (Gy)	75.85/75.05	59.90/63.25
Min (Gy)	66.45/68.55	44.35/45.85
Coverage	**0.948/0.947**	**0.976/0.965**
Homogeneity	**1.26/1.18**	
Conformity	**1.13/1.22**	

Table 22: DVH statistics of critical structures and normal tissues at Head-and-Neck case

	Spinal cord CG-P/FB-P	Brainstem CG-P/FB-P
Max(Gy)	36.45/54.55	36.05/41.75
Mean(Gy)	3.62/8.48	6.21/13.26
Min(Gy)	0.05/0.05	0.05/0.05
V_{20}	0.027/0.137	0.051/0.241
V_{40}	0/0.029	0/0.002
V_{60}	0/0	0/0

	R. parotid CG-P/FB-P	L. parotid CG-P/FB-P	R. submand. CG-P/FB-P	Tissue CG-P/FB-P
Max(Gy)	51.05/67.85	68.95/77.85	54.95/71.85	87.05/78.85
Mean(Gy)	7.61/17.91	20.91/37.97	20.59/42.1	29.79/30.86
Min(Gy)	0.05/5.45	0.15/2.25	2.65/17.15	0.05/0.05
V_{20}	0.064/0.243	0.408/0.705	0.439/0.905	0.638/0.667
V_{40}	0.008/0.035	0.202/0.500	0.124/0.591	0.351/0.362
V_{60}	0/0.006	0.018/0.198	0/0.106	0.074/0.074

3.6 Conclusion

A mixed-integer quadratic programming approach was presented for the optimization problems of IMRT treatment planning. The proposed MIQP model allows one to optimize simultaneously the beam orientations over the entire space of the gantry angles and the beamlet intensity map. For computational efficiency, a restricted branch-and-price method is derived to solve these large-scale MIQP instances arising from clinical cases of prostate and head-and-neck cancer. Clinical experiments showed that the solutions from our MIQP are far superior than those plans obtained in by the current best-practice in the clinical environment.

Beyond drastically improving the computational time, the restricted branch-and-price method provides an additional bonus: every node consists of only a relatively small number of beam variables in the branch-and-bound framework. This implies that in the middle of the solution process, a feasible solution obtained can be clinically usable, and it can be readily obtained by a few branching steps of the beam variables that are in the active set.

REFERENCES

[1] Ahnesjö, A., "Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media," *Medical Physics*, vol. 16, pp. 577–592, 1989.

[2] Ahnesjö, A. and Aspradakis, M., "Dose calculations for external photon beams in radiotherapy," *Phys Med Biol*, vol. 44, pp. R99–R155, 1999.

[3] Ahnesjö, A., Saxner, M., and Trepp, A., "A pencil beam model for photon dose calculation," *Medical Physics*, vol. 19, no. 2, pp. 263–273, 1992.

[4] Andersen, K., "A modified schur-complement method for handling dense columns in interior-point methods for linear programming," *ACM Transactions on Mathematical Software*, vol. 22, no. 3, pp. 348 – 356, 1996.

[5] Bahr, G., Kereiakes, J., Horwitz, H., Finney, R., Galvin, J., and Goode, K., "The method of linear programming applied to radiation treatment planning," *Radiology*, vol. 91, pp. 686–693, 1968.

[6] Barnhart, C., Boland, N., Clarke, L., Johnson, E., Nemhauser, G., and Shenoi, R., "Flight string models for aircraft fleeting and routing," *Transportation Sci.*, vol. 32, no. 3, pp. 208–220, 1998.

[7] Barnhart, C., Johnson, E., Nemhauser, G., Savelsbergh, M., and Vance, P., "Branch-and-price: Column generation for huge integer programs," *Operations Research*, vol. 46, no. 3, pp. 316–329, 1998.

[8] Battista, J. and Sharpe, M., "True three-dimensional dose computations for megavoltage x-ray therapy: A role for the superposition principle," *Phys Eng Sci Med*, vol. 15, pp. 159–178, 1992.

[9] Bednarz, G., Michalski, D., Anne, P., and Valicenti, R., "Inverse treatment planning using volume-based objective functions," *Physics in Medicine and Biology*, vol. 49, pp. 2503–2514, June 2004.

[10] Bednarz, G., Michalski, D., Houser, C., Huq, M., Xiao, Y., Anne, P., and Galvin, J., "The use of mixed-integer programming for inverse treatment planning with pre-defined field segments," *Physics in Medicine and Biology*, vol. 47, pp. 2235–2245, July 2002.

[11] Bortfeld, T., "Optimized planning using physical objectives and constraints," *Semin Radiat Oncol*, vol. 9, pp. 20–34, Jan 1999.

[12] Bortfeld, T., Burkelbach, J., Boesecke, R., and Schlege, W., "Methods of image reconstruction from projections applied to conformation radiotherapy," *Physics in Medicine and Biology*, vol. 35, pp. 1423–1434, 1990.

[13] Bortfeld, T., K., J., M., G., J., K., and SB., J., "Effects of intra-fraction motion on imrt dose delivery: statistical analysis and simulation," *Phys Med Biol*, vol. 47, pp. 2203 – 2, Jul 2002.

[14] Bortfeld, T., Stein, J., and Presier, K., "Clinically relevant intensity modulation optimization using physical criteria," in *XII Internation conference on the use of computers in radiation therapy*, (Madison, WI).

[15] Boyer, A. and Mok, E., "A photon dose distribution model employing convolution calculations," *Medical Physics*, vol. 12, pp. 169–177, 1985.

[16] Brahme, A., "Development of radiation therapy optimization," *Acta Oncol*, vol. 39, no. 5, 2000.

[17] C-M., M., Mok, E., Kapur, A., and et al., "Clinical implementation of a monte carlo treatment planning system," *Medical Physics*, vol. 26, pp. 2133–2143, 1999.

[18] Carol, M., "Integrated 3-d conformal multivane intensity modulation delivery system for radiotherapy," in *Proceedings of the 11th International Conference on the Use of Computers in Radiation Therapy*, (Madison, WI), 1994.

[19] Chen, Y., Michalski, D., Houser, C., and Galvin, J., "A deterministic iterative least-squares algorithm for beam weight optimization in conformal radiotherapy," *Physics in Medicine and Biology*, vol. 47, pp. 1647–58, May 2002.

[20] Cho, P. and Phillips, M., "Reduction of computational dimensionality in inverse radiotherapy planning using sparse matrix operations," *Physics in Medicine and Biology*, vol. 6, pp. N117–N125, May 2001.

[21] Choi, I., Monma, C., and Shanno, D., "Further development of a primal-dual interior point method," *ORSA J. Comput.*, vol. 2, pp. 304–311, 1990.

[22] Chui, C. and Spirou, S., "Inverse planning algorithms for external beam radiation therapy," *Med. Dosim.*, vol. 26, pp. 189–197, 2001.

[23] Cooper, R., "A gradient method of optimizing external-beam radiotherapy treatment plans," *Radiology*, vol. 129, no. 1, pp. 235–43, 1978 Jul.

[24] Cotrutz, C. and Xing, L., "Using voxel-dependent importance factors for interactive dvh-based dose optimization," *Physics in Medicine and Biology*, vol. 47, pp. 1659–69, May 2002.

[25] Crooks, S. and Xing, L., "Linear algebraic methods applied to intensity modulated radiation therapy," *Physics in Medicine and Biology*, vol. 46, pp. 2587–606, Oct 2001.

[26] Dantzig, G. and Wolfe, P., "Decomposition principle for dynamic programs," *Operations Research*, vol. 8, no. 1, pp. 101–111, 1960.

[27] Das, S. and Marks, L., "Selection of coplanar and noncoplanar beams using three-dimensional optimization based on maximum beam separation and minimized nontarget irradiation," *International Journal of Radiation Oncology Biology Physics*, vol. 38, no. 3, pp. 643–655, 1997.

[28] De Wagter, C., Colle, C., Fortan, L., Van Duyse, B., Van den Berge, D., and De Neve, W., "3d conformal intensity-modulated radiotherapy planning: interactive optimization by constrained matrix inversion," *Radiother Oncol*, vol. 47, pp. 69–76, Apr 1998.

[29] Deasy, J., Lee, E., Bortfeld, T., Langer, M., Zakarian, K., Alaly, J., Zhang, Y., Liu, H., Mohan, R., Ahuja, R., Pollack, A., Purdy, J., and Rardin, R., "A collaboratory for radiation therapy treatment planning optimization research," *Annals of Operations Research*. accepted Mar 2005.

[30] Deasy, J., "Multiple local minima in radiotherapy optimization problems with dose-volume constraints," *Medical Physics*, vol. 24, no. 7, pp. 1157–1161, 1997.

[31] Deasy, J., Blanco, A., and Clark, V., "Cerr: A computational environment for radiotherapy research," *Medical Physics*, vol. 30, pp. 979–985, 2003.

[32] Desrochers, M., Desrosiers, J., and Solomon, M., "A new optimization algorithm for the vehicle routing problem with time windows," *Oper. Res.*, vol. 40, no. 2, pp. 342–354, 1992.

[33] Desrosiers, J., Soumis, F., and Desrochers, M., "Routing with time windows by column generation," *Networks*, vol. 14, pp. 545–565, 1984.

[34] Djajaputra, D., Wu, Q., Wu, Y., and Mohan, R., "Algorithm and performance of a clinical imrt beam-angle optimization system," *Physics in Medicine and Biology*, vol. 48, pp. 3191–3212, 2003.

[35] D'Souza, W., Meyer, R., and Shi, L., "Selection of beam orientations in intensity-modulated radiation therapy using single-beam indices and integer programming," *Physics in Medicine and Biology*, vol. 49, pp. 3465–3481, Aug. 2004.

[36] Fraass, B., Kessler, M., McShan, D., Marsh, L., Watson, B., Dusseau, W., Eisbruch, A., Sandler, H., and Lichter, A., "Optimization and clinical use of multisegment intensity-modulated radiation therapy for high-dose conformal therapy," *Semin Radiat Oncol.*, vol. 9, no. 1, pp. 60–77, 1999 Jan.

[37] Gilmore, P. and Gomory, R., "A linear programming approach to the cutting-stock problem," *Operations Research*, vol. 9, pp. 849–859, 1961.

[38] Gilmore, P. and Gomory, R., "A linear programming approach to the cutting stock problem-part ii," *Operations Research*, vol. 11, pp. 863–888, 1963.

[39] Group, I. M. R. T. C. W., "Intensity-modulated radiotherapy: Current status and issues of interest," *International Journal of Radiation Oncology Biology Physics*, vol. 51, no. 4, pp. 880–914, 2001.

[40] Hartmann, S. C., Bergstrom, P., Chandler, W., and et al., "Lawrence livermore national laboratorys peregrine project," in *Proceedings of the XII International Conference on the Use of Computers in Radiation Therapy*, 1997.

[41] Hilbig, M., Hanne, R., Kneschaurek, P., Zimmermann, F., and Schweikard, A., "Design of an inverse planning system for radiotherapy using linear optimization," *Med Phys.*, vol. 12, no. 2, pp. 89–96, 2002.

[42] Holmes, T. and Mackie, T., "A comparison of three inverse treatment planning algorithms," *Phys Med Biol*, vol. 39, pp. 91–106, 1994.

[43] Hristov, D. and Fallone, B., "An active set algorithm for treatment planning optimization," *Med Phys.*, vol. 24, no. 9, pp. 1455–64, 1997 Sep.

[44] Johnson, E., "Modelling and strong linear programs for mixed integer programming," in *Algorithms and Model Formulations in Mathematical Programming* (Wallace, S., ed.), pp. 1–43, Berlin: Springer, 1989.

[45] Johnson, E., Mehrotra, A., and Nemhauser, G., "Min-cut clustering," *Math. Programming*, vol. 62, pp. 133– 151, 1993.

[46] Kalinin, E. and Deasy, J., "A method for fast 3-d imrt dose calculations: The quadrant infinite beam (qib) algorithm," *Medical Physics*, vol. 30, no. 6, pp. 1348–1349, 2003.

[47] Karmarkar, N., "A new plynomial-time algorithm for linear programming," *Combinatorica*, vol. 4, pp. 373–395, December 1984.

[48] Khan, F., *The Physics of Radiation Therapy. 2nd ed.* Baltimore: Williams and Wilkins, 1992.

[49] Kirkpatrick, S., Gelatt, C., and Vecchi, M., "Optimization by simulated annealing," *Science*, vol. 220, pp. 671–80, 1983.

[50] Kooy, H., Nedzi, L., Loeffler, J., Alexander, E., Cheng, C., Mannarino, E., Holupka, E., and Siddon, R., "Treatment planning for stereotactic radiosurgery of intracranial lesions," *International Journal of Radiation Oncology Biology Physics*, vol. 21, pp. 683–693, 1991.

[51] Langer, M., Brown, R., Urie, M., Leong, J., Stracher, M., and Shapiro, J., "Large scale optimization of beam weights under dose-volume restrictions," *International Journal of Radiation Oncology Biology Physics*, vol. 18, pp. 887–893, 1990.

[52] Langer, M., Morrill, S., Brown, R., Lee, O., and Lane, R., "A comparison of mixed integer programming and fast simulated annealing for optimizing beam weights in radiation therapy," *Medical Physics*, vol. 23, no. 6, pp. 957–964, 1996.

[53] Lee, E., "Computational experience with a general purpose mixed 0/1 integer programming solver (mipsol)," tech. rep., School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 1997.

[54] Lee, E., Fox, T., and Crocker, I., "Optimization of radiosurgery treatment planning via mixed integer programming," *Medical Physics*, vol. 27, no. 5, pp. 995–1004, 2000.

[55] Lee, E., Fox, T., and Crocker, I., "Integer programming applied to intensity-modulated radiation therapy treatment planning," *Annals of Operations Research, Optimization in Medicine*, vol. 119, pp. 165–181, 2003.

[56] Lee, E., Fox, T., and Crocker, I., "Simultaneous beam geometry and intensity map optimization in intensity-modulated radiation therapy," *International Journal of Radiation Oncology, Biology and Physics*, vol. 64, no. 1, pp. 301–320, 2006.

[57] Lee, E., Fox, T., and Crocker, I., "Simultaneous beam geometry and intensity map optimization in intensity-modulated radiation therapy," *International Journal of Radiation Oncology, Biology and Physics*, vol. 64, no. 1, pp. 301–320, 2006.

[58] Lee, E. and Mitchell, J., "Computational experience of an interior-point sqp algorithm in a parallel branch-and-bound framework," in *High Performance Optimization Techniques* (Franks, Roos, Terlaky, and Zhang, eds.), pp. 329–347, Kluwer Academic Publishers, 1997.

[59] Lee, E. and Zaider, M., "Mixed integer programming approaches to treatment planning for brachytherapy – application to permanent prostate implants," *Annals of Operations Research – Optimization in Medicine*, vol. 119, pp. 147–163, 2003.

[60] Llacer, J., Deasy, J., Bortfeld, T., Solberg, T., and Promberger, C., "Absence of multiple local minima effects in intensity modulated optimization with dose-volume constraints," *Physics in Medicine and Biology*, vol. 48, pp. 183–210, 2003.

[61] Lustig, I., Marsten, R., and Shanno, D., "On implementing mehrotra's predictor-corrector interior point method for linear programming," *SIAM Journal on Optimization*, vol. 2, no. 3, pp. 435–449, 1992.

[62] Lustig, I., Marsten, R., and Shanno, D., "Computational experience with a globally convergent primal-dual predictor-corrector algorithm for linear programming," *Mathematical Programming*, vol. 66, pp. 123–135, 1994.

[63] Lustig, I., Marsten, R., and Shanno, D., "Interior point methods for linear programming: Computational state of the art," *ORSA Journal on Computing*, vol. 6, no. 1, pp. 1–15, 1994.

[64] Lyman, J., "Complication probability as assessed from dose volume histograms," *Radiat Res*, vol. 104, pp. S13–S19, 1985.

[65] Lyman, J. and Wolbarst, A., "Optimization of radiation therapy. iii. a method of assessing complication probabilities from dose-volume histograms," *Int J Radiat Oncol Biol Phys*, vol. 13, pp. 103–109, 1987.

[66] Mackie, T., Reckwerdt, P., McNutt, T., and et al., "Photon beam dose computations," in *J. Palta, T.R. Mackie, editors. Teletherapy: Present and future*, (College Park, MD), Advanced Medical Publishing, 1996. p. 103-136.

[67] Mackie, T., Scrimger, J., and Battista, J., "A convolution method of calculating dose for 15 mv x rays," *Medical Physics*, vol. 12, pp. 188–196, 1985.

[68] Mageras, G. and Mohan, R., "Application of fast simulated annealing to optimization of conformal radiation treatments," *Medical Physics*, vol. 20, no. 3, pp. 639–647, 1993.

[69] Mehrotra, S., "On the implementation of a primal-dual interior point method," *SIAM Journal on Optimization*, vol. 2, no. 4, pp. 575–601, 1992.

[70] Mohan, R., Chui, C., and Lidofsky, L., "Differential pencil beam dose computation model for photons," *Medical Physics*, vol. 13, pp. 64–73, 1986.

[71] Mohan, R., Mageras, G., Baldwin, B., and et al., "Clinically relevant optimization of 3-d conformal treatments," *Medical Physics*, vol. 19, pp. 933–944, 1992.

[72] Mohan, R., Wang, X., Jackson, A., Bortfeld, T., Boyer, A., Kutcher, G., Leibel, S., Fuks, Z., and Ling, C., "The potential and limitation of the inverse radiotherapy technique," *Radiother. Oncol.*, vol. 32, pp. 232–248, 1994.

[73] Moller, M., "A scaled conjugate-gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, pp. 525–533, 1993.

[74] Morgan, W., Naqvi, S., Yu, C., Smith, L., and Bose, M., "Does the time required to deliver imrt reduce its biological effectiveness?," *International Journal of Radiation Oncology Biology Physics*, vol. 54, p. 22, 2002.

[75] Morrill, S., Lam, K., Lane, R., Langer, M., and Rosen, I., "Very fast simulated reannealing in radiation therapy treatment plan optimization," *International Journal of Radiation Oncology Biology Physics*, vol. 31, pp. 179–188, Jan 1995.

[76] Morrill, S., Lane, R., Jacobson, G., and Rosen, I., "Treatment planning optimization using constrained simulated annealing," *Physics in Medicine and Biology*, vol. 36, pp. 1341–1361, Oct 1991.

[77] Niemierko, A., "Reporting and analyzing dose distributions: A concept of equivalent uniform dose," *Med Phys*, vol. 24, pp. 103–110, 1997.

[78] Niemierko, A. and Goitein, M., "Calculation of normal tissue complication probability and dose-volume histogram reduction schemes for tissues with a critical element architecture," *Radiother Oncol*, vol. 20, pp. 166–176, 1991.

[79] Niemierko, A. and Goitein, M., "Modeling of normal tissue response to radiation: The critical volume model," *Int J Radiat Oncol Biol Phys*, vol. 25, pp. 135–145, 1992.

[80] Niemierko, A. and Goitein, M., "Implementation of a model for estimating tumor control probability for an inhomogeneously irradiated tumor," *Radiother Oncol*, vol. 29, pp. 140–147, 1993.

[81] Niemierko, A., Urie, M., and Goitein, M., "Optimization of 3d radiation therapy with both physical and biological end points and constraints," *Int J Radiat Oncol Biol Phys*, vol. 23, pp. 99–108, 1992.

[82] Nocedal, J. and Wright, S., *Numerical Optimization, Springer Series in Operations Research*. New York: Springer, 1999.

[83] Preciado-Walters, F., Rardin, R., Langer, M., and Thai, V., "A coupled column generation, mixed integer approach to optimal planning of intensity modulated radiation therapy for cancer," *Math. Program., Ser. B.*, vol. 101, pp. 319–338, 2004.

[84] Press, W., Vetterling, W., and Flannery, B., *Neumerical Recipes in C: The Art of Scientific Computing, 2nd ed.* Cambridge: Cambridge University Pree, 1992.

[85] Rockafellar, R. and Uryasev, S., "Conditional value-at-risk for general loss distributions," *J. Banking Finance*, vol. 26, pp. 1443–71, May 2002.

[86] Romeijn, H., Ahuja, R., Dempsey, J., and et al., "A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planning," *Physics in Medicine and Biology*, vol. 48, no. 21, pp. 3521–3542, 2003.

[87] Rosen, I., Lane, R., Morrill, S., and Belli, J., "Treatment plan optimization using linear programming," *Physics in Medicine and Biology*, vol. 18, pp. 141–152, 1991.

[88] Saunders, M., "Major cholesky would feel proud," *ORSA Journal on Computing*, vol. 6, no. 1, pp. 23–27, 1994.

[89] Savelsbergh, M., "A branch-and-price algorithm for the generalized assignment problem," *Oper. Res.*, vol. 45, no. 6, pp. 831–841, 1997.

[90] Schach von Wittenau, A., Cox, L., Bergstrom, P., and et al., "Correlated histogram representation of monte carlo derived medical accelerator photon-output phase space," *Medical Physics*, vol. 26, pp. 1196–1211, 1999.

[91] Sempau, J., SJ., W., and Bielajew, A., "Dpm, a fast, accurate monte carlo code optimized for photon and electron radiotherapy treatment planning dose calculations," *Phys Med Biol*, vol. 45, pp. 2263–2291, 2000.

[92] Shepard, D. M., Ferris, M. C., Olivera, G. H., and Mackie, T. R., "Optimizing the delivery of radiation therapy to cancer patients," *SIAM Rev.*, vol. 41, pp. 721–744, 1999.

[93] Smith, A., Ling, C., and eds, "Implementation of three dimensional conformal radiotherapy," *Int J Radiat Oncol Biol Phys.*, vol. 33, pp. 779–976, 1995.

[94] Smith, A., Purdy, J., and eds, "Three-dimensional photon treatment planning. report of the collaborative working group on the evaluation of treatment planning for external photon beam radiotherapy," *Int J Radiat Oncol Biol Phys*, vol. 21, pp. 3–268, 1991.

[95] Society, A. C., *Source: Cancer Facts and Figures – 2006.* 2006.

[96] Spirou, S. and Chui, C., "A gradient inverse planning algorithm with dose-volume constraints," *Medical Physics*, vol. 25, no. 3, pp. 321–333, 1998.

[97] Starkschall, G., "A constrained least-squares optimization method for external beam radiation therapy treatment planning," *Med Phys.*, vol. 11, pp. 659–665, 1984.

[98] Stein, J., Mohan, R., Wang, X., Bortfeld, T., Wu, Q., Preiser, K., Ling, C., and Schlegel, W., "Number and orientations of beams in intensity-modulated radiation treatments," *Medical Physics*, vol. 24, pp. 149–160, Feb. 1997.

[99] Szu, H. and Hartley, R., "Fast simulated annealing," *Phys. Lett. A.*, vol. 122, pp. 157–162, 1987.

[100] Treuer, U., Treuer, H., Hoevels, M., Muller, P., and Sturm, V., "Computerized optimization of multiple isocenters in stereotactic convergent beam irradiation," *Physics in Medicine and Biology*, vol. 43, pp. 49–64, 1998.

[101] van den Akker, J., Hurkens, C., and Salvesbergh, M., "A time-indexed formulation for single-machine scheduling problems: Column generation," *INFORMS Journal on Computing*, vol. 12, no. 2, pp. 111–124, 2000.

[102] Vance, P., Barnhart, C., Johnson, E., and Nemhauser, G., "Solving binary cutting stock problems by column generation and branch-and-bound," *Comput. Optim. Appl.*, vol. 3, no. 2, pp. 111–130, 1994.

[103] Vanderbeck, F., "Computational study of a column generation algorithm for bin packing and cutting stock problems," *Math. Programming*, vol. 86, no. 3, pp. 565–594, 1999.

[104] Vanderbeck, F., "On dantzig-wolfe decomposition in integer programming and ways to perform branching in a branch-and-price algorithm," *Oper. Res.*, vol. 48, no. 1, pp. 111–128, 2000.

[105] Vanderbeck, F. and Wolsey, L., "An exact algorithm for ip column generation," *Oper. Res. Lett.*, vol. 19, pp. 151–159, 1996.

[106] Wang, J., Li, X., DSouza, W., and Stewart, R., "Impact of prolonged fraction delivery times on tumour control probability: a note of caution for imrt," *International Journal of Radiation Oncology Biology Physics*, vol. 57, p. 543552, 2003.

[107] Webb, S., "Optimisation of conformal radiotherapy dose distributions by simulated annealing," *Physics in Medicine and Biology*, vol. 34, pp. 1349–1370, 1989.

[108] Webb, S., "Optimization by simulated annealing of three-dimensional conformal treatment planning for radiation fields defined by a multileaf collimator," *Physics in Medicine and Biology*, vol. 36, pp. 1201–1226, 1991.

[109] Webb, S., *The Physics of Three-Dimensional Radiation Therapy: Conformal Radiotherapy, Radiosurgery and Treatment Planning.* Philadelphia: Institute of Physics Publishing, 1993.

[110] Webb, S., *Intensity Modulated Radiation Therapy.* Bristol: Institute of Physics Publishing, 2001.

[111] Woo, M., Cunningham, J., and Jerioranski, J., "Extending the concept of primary and scatter separation to the condition of electronic disequilibrium," *Medical Physics*, vol. 17, pp. 588–595, 1990.

[112] Wu, C., Jeraj, R., Olivera, G., and Mackie, T., "Re-optimization in adaptive radiotherapy," *Physics in Medicine and Biology*, vol. 47, pp. 3181–3195, Sept. 2002.

[113] Wu, Q. and Mohan, R., "Algorithms and functionality of an intensity modulated radiotherapy optimization system," *Medical Physics*, vol. 27, no. 4, pp. 701–711, 2000.

[114] Wu, Q. and Mohan, R., "Multiple local minima in imrt optimization based on dosevolume criteria," *Medical Physics*, vol. 29, pp. 1514–1527, 2002.

[115] Xing, L., Li, J., Donaldson, S., Le, Q., and Boyer, A., "Optimization of importance factors in inverse planning," *Physics in Medicine and Biology*, vol. 44, pp. 2525–2536, 1999.

[116] Yang, Y. and Xing, L., "Inverse treatment planning with adaptively evolving voxel-dependent penalty scheme," *Medical Physics*, vol. 31, no. 10, pp. 2839–2844, 2004.

[117] Yu, Y., Schell, M., and Zhang, J., "Decision theoretic steering and genetic algorithm optimization: Application to stereotactic radiosurgery treatment planning," *Medical Physics*, vol. 24, pp. 1742–1750, 1997.

[118] Zhang, X., Liu, H., Wang, X., and Dong, L., "Speed and convergence properties of gradient algorithms for optimization of imrt," *Medical Physics*, vol. 31, pp. 1141–1152, May 2004.

VITA

Kyungduck Cha received Bachelor of Science and Master of Science in Industrial Engineering from Gyeongsang National University, Chinju, Korea in 1997 and 1999, respectively. He joined the Ph.D. program in the School of Industrial and Systems Engineering at the Georgia Institute of Technology in 2000 where he worked with Dr. Eva Lee as research assistant at the Center for Operations Research in Medicine and HealthCare. His research interests are in mathematical programming and large-scale computational algorithms. His recent works are involved in bio/medical decision systems supporting system modelling, algorithm and software design, and decision theory analysis.